Citation for published version:
Haynes, D, Streatfield, D, Jowett, T & Blake, M 1997, *Responsibility for Digital Archiving and Long Term Access to Digital Data*. JISC/NPO Studies on the Preservation of Electronic Materials, British Library Research and Innovation Centre, London.

Publication date:
1997

Document Version
Publisher's PDF, also known as Version of record

Link to publication

University of Bath

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

British Library Research and Innovation Report 67

JISC/NPO STUDIES ON THE PRESERVATION OF ELECTRONIC MATERIALS

Responsibility for digital archiving and long term access to digital data

David Haynes, David Streatfield, Tanya Jowett and Monica Blake

British Library Research and innovation Centre
1997

This study is part of a programme funded by JISC as a result of a workshop on the Long Term Preservation of Electronic Materials held at Warwick in November 1995.

The programme of studies is guided by the Digital Archiving Working Group, which reports to the Management Committee of the National Preservation Office.

The programme is administered by the British Library Research and Innovation Centre.

Abstract

This report presents the findings of an investigation into opinions on the responsibility for maintaining an archive of digital materials produced in the UK and Ireland. The study was conducted by means of focus group meetings and interviews. The report represents the range of opinions expressed and highlights areas of concern to the participants, all of whom have an interest in this issue.

The report recommended that a body should be established to co-ordinate archiving of digital material. The National Office of Digital Archiving (NODA) would be responsible for developing standards and guidelines for archiving digital materials. The job of maintaining the archives should be contracted out to specialist agencies with the appropriate expertise. Once material is selected for preservation it should be kept for ever. Funding for digital archiving must come from the public sector via higher education institutions, legal deposit libraries, the funding councils and NODA itself. Legal deposit legislation should be amended to cover electronic publications and other digital material such as sound and video recordings.

Authors

This project was jointly conducted by David Haynes of David Haynes Associates and David Streatfield of Information Management Associates with input from Tanya Jowett, a consultant with David Haynes Associates, and Monica Blake an independent consultant.

© Joint Information Systems Committee of the Higher Education Funding Councils 1997

The opinions expressed in this report are those of the authors and not necessarily those of the sponsoring bodies or the British Library

RIC/CT/305

ISBN 0 7123 33258

ISSN 1366-8218

British Library Research and Innovation Reports may be purchased as a photocopy or microfiche from the British Thesis Service, British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, United Kingdom. Tel: 01937 546229; Fax: 01937 546286; email dsc-british-thesis-service@bl.uk

To borrow a copy contact: DSC Customer Service, The British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, United Kingdom. Tel: 01937 546060; Fax: 01937 546333; email: dsc-customer-services@bl.uk

Contents

INTRODUCTION

Background

In 1995, FIGIT and BLR&DD (British Library Research and Development Department) co-sponsored a workshop on the Long Term Preservation of Electronic Materials at Warwick University. The workshop resulted in a list of actions, which can be found in the report of the workshop, prepared by Marc Fresko, at:

http://www.ukoln,ac.uk/resko/warwick/intro.html

The list of actions was considered by the Management Committee of the National Preservation Office, and JISC subsequently agreed to fund a programme of studies, which has been developed in conjunction with the National Preservation Office and administered by the British Library Research and Innovation Centre (BLRIC). The programme of studies is guided by the National Preservation Office Digital Archiving Working Group, chaired by Peter Fox, Librarian of Cambridge University who is a member of the National Preservation Office Management Committee.

The following topics are the subject of projects to be completed in 1997:

1. Analysis of the report from the Research Libraries Group/Commission on Preservation and Access for its relevance and applicability in the UK
2. Framework of major data types and formats identifying issues affecting preservation of each category of material
3. An investigation of the attitudes of originators and rights' holders to the responsibilities of digital preservation
4. A study of the three main methods of digital preservation: technology preservation; technology emulation; information migration
5. An investigation into the digital preservation needs of universities and research funders
6. An investigation of progress already made towards permissive guidelines for digital preservation
7. Report on sampling methods and techniques for collecting materials, on the nature and extent of institutional electronic archives, and the relevance of current archival practice to digital preservation

This report summarises the findings of the third project in this list:

An investigation of the attitudes of originators and rights' holders to the responsibilities of digital preservation

Terms of Reference

The aims of this project were:

To investigate the attitudes of originators and rights owners to their responsibilities for the preservation of digital data.

The consultants were specifically tasked to:

interview a range of publishers, universities, libraries with special collections, research organisations, data compilers and repositories to obtain knowledge of the policies adopted across a range of institutions particularly with regard to:

willingness/perceived ability to preserve/hold data in perpetuity

willingness/desire to pass on this responsibility

ability to preserve unpublished data where accompanying documentation/bibliographic details are insufficient

Models to be tested include:

publisher retains database

database deposited with a research library/digital archive consortium.

Approach

There have been a number of initiatives to consider some of the issues surrounding the preservation of electronic materials. The British Library/JISC workshop on the Long Term Preservation of Electronic Materials identified many of the issues that need to be considered in the development of a digital archive. Many of the stakeholders were identified including:

Libraries
Publishers
Archive centres
Distributors
IT suppliers
Legal depositories
Consortia

We would add to these Authors as distinct from Publishers, and Networked information service providers as distinct from Distributors.

The purpose of this project was to identify the issues that will influence the decisions on where the responsibility should lie for keeping digital archives and how they are funded. We were keen that in the first instance all the different interest groups were identified before the consultation exercise began. Our approach was to consider the publication cycle which applies to electronic products and identify the key players at each stage in the process. This model provided us with the basis for selection of organisations and individuals for interview. Some individuals may represent more than one interest group and this was taken into account in identifying the relevant individuals.

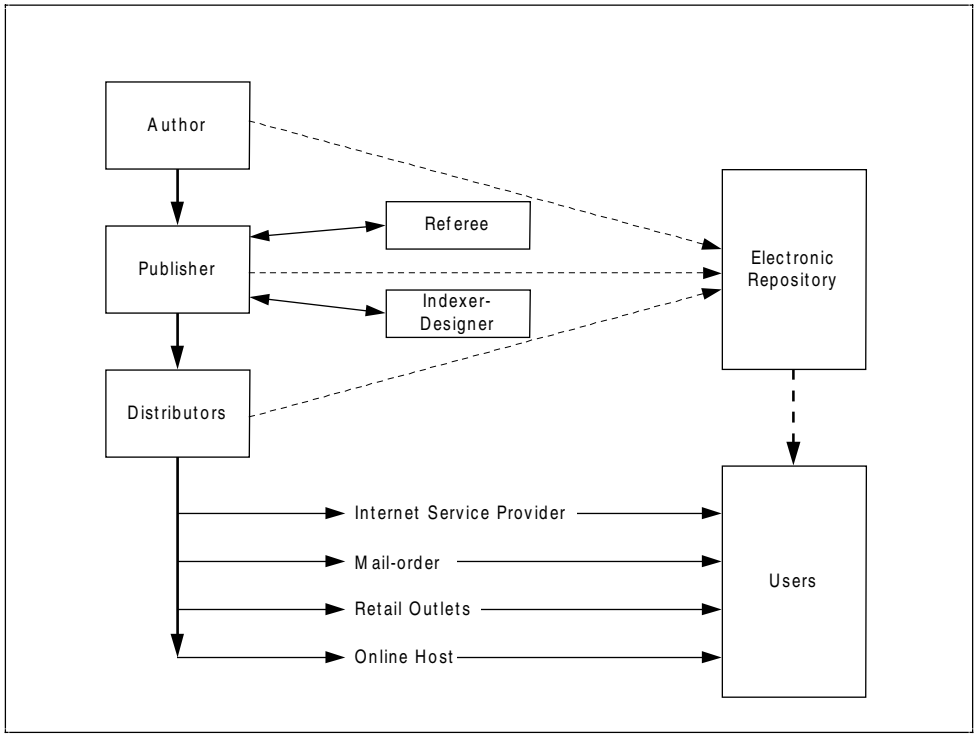

Figure 1 - Production of Electronic Publications

Method

The first phase of the project was to conduct literature searches to identify recent developments in the area of digital archiving internationally. This review also provided a means of identifying potential participants in the focus groups, and names of individuals for interview.

Develop models

The project documents included two outline models for digital archiving. These provided the basis for discussions with different groups and were used to develop more detailed descriptions of the options available for archiving of data.

Focus groups

A series of focus group meetings was arranged, based around the following interest groups:

Authors & research funders

Publishers

Distributors

Repositories

See Appendix B for the accounts of these focus groups.

Interviews

There are many different interests that need to be represented in an initiative of this kind. We conducted 39 face-to-face and telephone interviews (see Appendix A for

a list of those consulted) with representatives of the major stake holding groups. The interviews provided an opportunity to test the models proposed for the archiving of data and clarify some of the issues raised during the seminar/focus group meetings. The questionnaire used as the basis of the telephone and face-to-face interviews is in Appendix C.

LITERATURE REVIEW

Background

RLG/CPA Report

Much of the current debate on digital archiving was started by the Task Force on Digital Archiving, a group which was created in the USA by the Commission on Preservation and Access and The Research Libraries Group. The report of that Task Force (11) has been widely quoted, criticised and praised. The purpose of the Task Force was to investigate the means of ensuring continued access indefinitely into the future of records stored in digital electronic form. Composed of individuals drawn from industry, museums, archives and libraries, publishers, scholarly societies and government, the Task Force was charged specifically to:

> *"Frame the key problems (organisational, technological, legal, economic etc.) that need to be resolved for technology refreshing to be considered an acceptable approach to ensuring continuing access to electronic digital records indefinitely into the future.*
>
> *Define the critical issues that inhibit resolution of each identified problem.*
>
> *For each issue, recommend actions to remove the issue from the list.*
>
> *Consider alternatives to technology refreshing.*
>
> *Make other generic recommendations as appropriate"*

For the purposes of this study, the key section in the Task Force report is "Archival Roles and Responsibilities". The Task Force recommends the development of a national system of digital archives which would act as repositories of digital information. The repositories would be held together in a national archival system with two essential mechanisms in place:

- repositories claiming to serve an archival function must be able to prove that they are who they say they are by meeting or exceeding the standards and criteria of an independently-administered program for archival certification;
- certified digital archives will have available to them a critical fail-safe mechanism, which would be a legal right. This would enable them to exercise an aggressive rescue function to save culturally significant digital information.

The Task Force felt that:

> *Without the operation of a formal certification program and a fail-safe mechanism, preservation of the nation's cultural heritage in digital form will likely be overly dependent on marketplace forces, which may value information for too short a period and without applying broader, public interest criteria.*

They also believe that responsibility for archiving rests initially with the creator or owner of the information.

The national system of digital archives would have a distributed, rather than centralised, structure for collecting digital information objects.

> *A distributed structure, built on a foundation of electronic networks, places archival responsibility with those who presumably care most about and have the greatest understanding of the value of particular digital information objects. Moreover, such a structure locates the economic and cultural incentives where they are most likely to prompt those preserving digital information to respond with the greatest agility to the changing digital landscape and to the shifting tides of technology.*

The report makes nine recommendations including: securing proposals from potential digital archives, securing funding, encouraging experiments and demonstration projects, encouraging legislation on fail-safe mechanisms, and encouraging dialogue on standards for certifying repositories.

The report only briefly mentions the idea of legal deposit, which is being pursued in the UK, suggesting that it is another means of providing a fail-safe mechanism. It is unclear however, how you can ensure that originators preserve their material, or make partnerships to preserve it, without a legal deposit system. There could be many thousands of originators and it would be difficult to ascertain whether they are all preserving their material properly. Although it is proposed in the report that archives are given rights for "*aggressive rescue*", policing such a system could be very difficult.

Warwick Workshop

In November 1995, shortly after the publication of the draft CPA/RLG Report, a workshop was held at Warwick University to discuss how these issues affected the UK. The aims of the workshop were to:

- explore strategic options for developing and managing electronic archives;
- consider possible collection policies for electronic materials;
- consider preservation policies;
- examine practical implications of the various options;
- propose research and action needed, to resolve/shed light on the issues discussed.

The report of the workshop (5) highlighted eighteen potential action points. The National Preservation Office (NPO) subsequently identified eight projects which will be commissioned in 1997 to answer some of those points. This study is one of those actions.

One of the syndicate discussions held after the main presentations suggested that publishers cannot be relied on to preserve their digital publications and that no single body can take on the task of digital preservation in its entirety, as it is too big. An investigation of the attitudes of rights' owners to the responsibilities of digital preservation was therefore requested.

Loughborough Report

Another action point highlighted by the Warwick Report (5) was the need to analyse the RLG/CPA Report (11) to ascertain its relevance and applicability in the UK. This has now been done by Loughborough University and their report (6) was produced in 1997. It examined the recommendations of the RLG/CPA Report and came up with eight prioritised actions which were agreed at a meeting held at the British Library in December 1996. These include:

- appointing a National Digital Preservation Officer,
- establishing a National Digital Preservation Body,
- investigating digital archive practice and policy in the UK,
- identifying good practice and gaps in knowledge,
- devising guidelines on practice and a digital preservation policy,
- raising awareness,
- promoting education and training,
- fostering international co-operation.

Some of these actions were in hand by the time the Loughborough report had been published or have been put in place since. The first meeting of the national lead body, the Digital Archiving Working Group, to focus on the work being carried on by libraries and archives in the United Kingdom and Ireland was held in January 1997.

The National Preservation Office under the leadership of its Director has developed an extensive digital remit and undertakes a leadership role on behalf of the library and archive community at local, national and international levels. It provides an independent focus for ensuring the preservation and accessibility of library and archive materials and, as such, acts as the National Digital Preservation Body mentioned in the Loughborough Report.

Legal Deposit

It is widely felt that the most efficient way to ensure that digital material is archived in the UK is to extend the legal deposit legislation so that it includes electronic publications. The British Library has been considering this idea for some years and have submitted a proposal to the Department of National Heritage suggesting how the law could be changed (10). It recommends:

- the national published archive be a distributed archive (i.e. it recommends that more than one repository be given the right to receive publications through legal deposit);
- new legislation should enable the comprehensive deposit of non-print publications (designated repositories to apply selection criteria if appropriate);
- new legislation should give a Minister of the Crown the authority to specify where particular categories of materials be deposited;
- all designated repositories should meet required standards of storage, maintenance, preservation, bibliographic control and public use facilities;
- a body independent of publishers and independent of the designated repositories should adjudicate over categories of items for which there is any doubt about their suitability for deposit;
- legal deposit should apply to publications currently deposited through voluntary agreements;
- new legislation should apply not only to publications traditionally associated with print on paper and now appearing in electronic format, but also to all current and future UK non-print publications, including films and sound recordings, microform publications and digital publications (both off-line and on-line);

- new primary legislation should not attempt to specify all the publishing media within its scope but that the statute should provide for subsidiary legislation, contained in regulations, to be drawn up to cover new publishing media as they appear;
- "publication" be defined to include items or copies of items made available to the public through sale, hire, distribution without charge, broadcast, performance before an audience open to the public, or a technology enabling the public to read, hear or otherwise use or consult it in whole or in part;
- there should be an incremental approach to the implementation of legal deposit for print publications. The proposal identifies four main categories of non-print publications available today: (i) texts in microform, (ii) texts in "hand-held" electronic form, (iii) films and sound recordings, (iv) on-line publications. It recognises on-line publications as potentially the most important means of scholarly communication in the long term. However, because little is known about how to deposit them and provide access to them on a controlled basis, it recommends that legislation be drawn up for the other three categories first;
- there should be two-copy deposit of each of sound recordings and films, and up to six-copy deposit of microform publications and hand-held electronic publications;
- there should be no change to copyright legislation. The proposal treats legal deposit as independently as possible from copyright;
- the existing legal deposit legislation for printed publications should continue to apply.

In a statement by the LA/JCC Working Party on Copyright (2) the need for legislation was endorsed by a consortium of various bodies representing the library and information professions.

In response to the British Library's proposal, the Department of National Heritage has issued a consultation paper (3) which requested comments about the deposit of digital publications as well as many other issues to do with legal deposit arrangements. The consultation period is now over, but the recent change of government means it is unlikely that new legislation will be introduced to Parliament before 1998.

Research

There is a great deal of research being carried out into various aspects of digital archiving and we have not attempted to look at all of it here. The National Preservation Office is overseeing the studies funded by the Joint Information Systems Committee (JISC) through the eLib Programme, which were identified as a result of the Warwick workshop.

In 1996 Cimtech Ltd conducted a study on the *Preservation of Digital Material* (4). It was commissioned, along with three other studies, by the British Library Working Group on the Legal Deposit of Non-Print Materials, in order to inform its work. This report considers the main options for the long-term preservation of digital publications and estimates the costs involved with each one. The costings are only estimates and are very broad, but they are the only ones we have available currently. The report summarises the activities of other national libraries, particularly those in Norway, USA, Netherlands, France and Canada.

The British Library's Research and Innovation Centre is leading the Library's Digital Library Programme which also commissions research into different ways of exploiting the Library's information resources. The International Institute for Electronic Library Research at De Montfort University is also doing much research in this area, as is the Department of Information Studies at Loughborough University.

Current Initiatives

There is much literature which describes current activities and initiatives in the UK and abroad. We tried to get a general impression of some of these activities and would like to draw attention to a few publications, mainly on the World Wide Web, which describe these and point to other useful sites.

The National Library of Canada conducted an Electronic Publications Pilot Project (EPPP) to identify and understand all the challenges associated with acquiring, cataloguing, preserving and providing access to Canadian electronic publications. The goals of the project were (among other things):

- to identify and understand the issues that libraries encounter when handling on-line collections.
- to help the NLC determine long-term policies on how to handle electronic publications and recommend which areas in the NLC should handle these documents.
- to use some of the technologies involved in electronic publishing.

The summary of the final report (9) provides useful experience on issues like selection, standard formats, access and copyright, HTML links, technical issues, legal deposit and storage.

The National Library of Australia has produced a *Statement of Principles on the Preservation of and Long-Term Access to Australian Digital Objects* (7). There are eight principles covering co-operation, the role of the creator, distributed responsibility, appraisal, rights, strategies and the role of government. Principle 2 on the role of the creator states that creators of digital objects have the initial, and in some cases a continuing, role in preserving access to them. Principle 3 on distributed responsibility says that the location, selection, identification, cataloguing and retention of digital objects will be best achieved through the co-ordinated distribution of responsibilities. A national network of archives, collecting organisations and other information providers operating through formal agreements in which responsibilities are well-defined will best serve to preserve access to significant digital objects.

The Arts and Humanities Data Service (AHDS) has a very useful web page which gives access to the pages of all its service providers. These service providers are other data archives, often covering specific subject areas, whose archives form part of the AHDS network. AHDS has also provided a Digital Preservation page (1) which includes hypertext links to initiatives and projects all over the world related to digital preservation.

FINDINGS FROM THE INTERVIEWS

A series of interviews was conducted with nominated individuals in a selection of organisations representing several categories of stakeholder including: rights holders, distributors and repository managers. See Appendix A for a list of

participants and Appendix C for the interview questionnaire. Most of the interviews took place by telephone, but some visits were arranged for face-to-face discussion, especially where it was felt to be necessary to consult more than one person.

This section gives details of some of the answers to the questions which we have grouped under the following headings:

- Organisational policy
- Responsibility for archiving
- Procedures for archiving
- Timescale
- Format of material
- Funding for archiving
- Legal and commercial considerations
- Other legal considerations

Organisational policy

Of the 32 organisations that responded to the question on whether they had a policy on digital archiving, 16 had a policy and 8 of the remaining organisations were working towards a policy or felt that a policy was necessary:

> *No. We do not have a policy at the moment. ...have talked it over but there is one needed. We also need a policy on our own archives which are a valuable resource.*
>
> *The strategy is to find out what we ought to do! Trying to get a commitment in relation to the content that we are creating depositing the material, but this is not a policy yet.*
>
> *...haven't seen one in print. We have a commitment to maintaining the record*

Interestingly, one organisation was already aware of the need to have an archiving strategy as a condition of research grants:

> *No. Working on trying to get digitisation policy off the ground. Will need a policy before we can get grants.*

There was a perception among some of the respondents that the responsibility for developing an archiving policy lies elsewhere:

> *There are international organisations which are developing guidelines.*
>
> *Refer to the Council for Scientific Unions, the US National Academy of Sciences and the All European Academy Group.*
>
> *Have to abide by Public Record Act.*

For those organisations that did have an archiving policy there was a focus on physical storage:

> *...kept a fireproof cabinet full of Winchester disks. We are concerned about what might have happened to the archive following staff changes and a move to new premises.*
>
> *The strategy involves a daily back-up of tapes, and a weekly back-up to an off-site store*
>
> *Material is archived on tapes. The company operates a UNIX network. From time to time, snapshots are taken of the network. Two copies of archived material are kept; one copy is kept off-site in fireproof conditions.*
>
> *We archive our own product information using standard computer techniques - daily / weekly etc. backups, copies on site, in secure vaults elsewhere etc.*

Some publishers and distributors have a comprehensive policy:

> *As a publisher, our concern is to preserve our own material. For this, there is a strict archiving policy. Material that is licensed from a third party is archived when it comes into the building and immediately prior to publication. It is kept in raw and built forms. All versions of software are archived including ones that are never published. Archiving includes type definitions and SGML data.*
>
> *Yes very much, for own benefit not for benefit of society. Enormous redundancy built into storage. Keep everything under the ... mountains. Much duplication. This is common throughout the corporate world*
>
> *Well worked out policies. Master plates for all issues of CD-ROMs are stored as well*
>
> *Yes preserve own material. Copies of all electronic publications going back to early days. Difficult to access. Keep a copy of the final format. It used to be difficult to copy a database because of the size; now cut latest snapshot. As soon as product is released that version is copied. Do not have any online publications but are considering some. Archive is stored in a warehouse.*

One publisher's policy was driven by the issue of access:

> *Those products we have in digital form, we archive in a way that is accessible to others.*

One of the data depositories also focused on user access:

> *Taking data of use for research and teaching purposes to promote informed use. Dissemination and preservation function. Archive and library functions. Mainly but not just academic sector. Wide range of services.*

It had a comprehensive policy:

> *Responsibility for preserving it. Processing, validation, supplementing, scanning, creating metadata in digital form. Take all these together with data, strip off software dependencies. Checks for corruption, write to three different media. Then there is a continuing programme of checking for corruption, deciding when to move to new media. Archive held at different sites.*

Responsibility for archiving

There was a very diverse range of views on who should be responsible for holding and archiving digital materials. The most frequently mentioned was a National Digital Repository (6) or variations on that theme:

> *Most secure way of preserving digital material is for responsibility to be assigned to **national repositories**.*
>
> *Need to rely on **national repositories**. Can't be held locally. Economies of scale. Staffing levels and technology needed nationally. Makes most sense to split the archive up by disciplines as long as they talk to each other...It needs a **national body** to pull it together and ensure there are no gaps and nothing duplicated.*
>
> *... favours a system of depositing a single copy with a single designated repository.*
>
> *...establishment of a **National Digital Preservation body** in which all the stakeholders are represented. It is important to have a national forum for all the different interests to get together and find out what each other is doing and what research is going on etc.*

They also suggested a National Digital Preservation Officer be appointed.

Although there was one clear warning that:

> *A **central depository is unlikely to fulfil the needs** of different depositors.*

There was a cluster of suggestions for national libraries (5), the British Library (4) and the legal deposit libraries (4) to take on the responsibility for digital archives:

> *Should be the role of the **national library**, as with printed material.*

*If material is to do with Wales, then **National Library of Wales** has an interest regardless of format.*

*Responsibility can't be left to rest with producers of digital material. Unreasonable to ask commercial producers to carry overheads for preservation. Has to be some form of national organisation, e.g. **national library** or **national body***

*The **British Library** should know where things are. They cannot be expected to hold everything (e.g. non-English language veterinary publications, which are held by the Royal College of Veterinary Surgeons), but they should be able to refer people to appropriate organisations.*

*The **British Library** expects to be a deposit library for archiving digital materials, but the question is what is the scope of the material that will be kept. It needs to consider a selection policy of digital materials as well as for hardcopy. The **national library** should take on more responsibility but needs resources to do this.*

*Got to have an **independent national institution**. The **British Library** or an agency. ...concerned that they are too closely modelling the thinking on existing deposit libraries. You only need one electronic deposit which can give access to anyone, or a maximum of two to mirror for safety. Flexing of muscles by deposit libraries to retain control is a bit silly.*

*Creators of digital material should act responsibly. They should not set out to create material unless they think through the long-term implications. Suspect that in practice **deposit librarians** will have to take responsibility for the preservation of digital material - possibly with the higher education funding councils (now that HE bodies create a great deal of digital material).*

*The present situation is that once material is of less potential value it is made more widely available. The **legal deposit library** structure provides a basis for digital archiving policy. Whether or not the material is physically stored in the legal deposit libraries (LDLs) is an open question, although storing in just one location is too risky. Could be stored in several locations linked to the LDLs.*

*Our preference would be that it all stays with the **legal deposit libraries** that currently exist and should not be shared out amongst lots of archives.*

*Obvious answer is the **copyright libraries**. Not just the British Library. The British Library makes the thing remote. As far as Scotland is concerned, archive material should be the responsibility of **National Library of Scotland.***

The National Sound Archive was mentioned a number of times:

*For sound recordings ...favour continuing the existing voluntary system of deposit with the **National Sound Archive**.*

*The **National Sound Archive** should have responsibility for sound recordings.*

*The record industry supports the **National Sound Archive** and most companies deposit with it voluntarily. However it is accepted that there could be gaps in the sound archive using this system and that they should be eliminated. We support the extension of legal deposit to cover sound recordings on the condition that there is adequate protection for the owners of copyright. Neither the means of supply nor the penalties for non-compliance should be burdensome. Cost of compliance is great for small companies so publishers of sound recordings should only have to deposit one copy to the National Sound Archive. CDs are not as vulnerable as vinyl records so it is not necessary to keep a second copy*

Another popular suggestion was for a system of distributed depositories or a body to co-ordinate different depository agencies (5):

*In the case of the refereed learned journal literature, I think there should be an **international consortium** of Universities and learned societies, with governmental support, overseeing the archiving and preservation. As more and more of the literature does accumulate on the Internet, these higher-order alliances will coalesce; currently the electronic corpus is still too sparse and fragmented for a concerted preservation initiative: First we need an international, interdisciplinary corpus of at least a certain critical mass. Once the intellectual goods are really committed to the medium in earnest - not just thinking about it while contemplating preservation - then the interests vested in the corpus will drive the measures taken to preserve it.*

*Object to the DNH discussion document because it only referred to the British Library. It would be desirable to have multiple sites. There could be a **network of data archives**, each funded separately and in different ways.*

*We would need some kind of **co-ordinating body** to decide where the individual material went. This body could also decide on issues like saving snapshots of things every month etc.*

*Divided between local responsibility and centralised responsibility. Need a **national policy** as libraries are hard pressed*

*Responsibility has to be accepted by a **partnership** between national libraries, some special libraries (broadly defined to include organisations like the PRO) and the HE community. There should be a network of organisations which take responsibility for preservation.*

***Network of organisations**. Central policy, but carried out by lots. One organisation could not cope. Has to be a distinction between public and private information. Can't have a policy about private information, but may want an advisory service and alert them that the information is part of our heritage and could get lost. Public information, how do you define this especially as government departments become more often Agencies. Must have a policy, but there may be different organisations with responsibility. Central role for setting policies and ensuring things don't fall through the gap. Need education role, advisory service some auditing function to make sure it gets done.*

*There should be a **network of organisations** which take responsibility for preservation. Not clear whether to restrict this to a small number. Certainly, for the bulk of published material preservation should be done through a small number of libraries. Every library in UK could take some responsibility for some material. Which material needs to be dealt with by a specialist library needs serious thought. Things will get missed, no point worrying excessively about that. Some papers get missed or destroyed inadvertently or on purpose which would have been useful. It will be very complicated to work out what needs preserving. Much of Usenet is like the chattering in the pub but other stuff on there is useful. People will have to spend time deciding what to preserve and different kinds of organisation will preserve different type of things. There is a range of possibilities for a range of aims.*

Some respondents were unable to be more specific than to say that digital archiving should be the responsibility of Government departments, the government or the public domain (4):

*Need collaboration of different bodies to ensure preservation of the widest variety of archives. To rely on marketplace actors is probably a mistake. Need a **public institution** with ultimate responsibility.*

*In the end it has to be a **government responsibility** because of resources and long-terms interests. Government includes the British Library.*

*Cynical answer is not the corporates. They would be always doing it for own gain, could be hidden agendas. Wouldn't want that responsibility themselves, nor even less think that their competitors had responsibility for archiving. Has to be in **public domain**. Corporates would support it financially.*

*Clearly the **Department of National Heritage** [now the Department of Media Culture and Sports] as it sets policies for library community.*

***Government departments** ought to be responsible for archiving their own records.*

Several publishers suggested that they or the rights holders should archive their own material (4):

Very difficult issue. The received wisdom seems to be that publishers don't want the expense and bother of archiving something forever. Thus they pass on material when it's of no more commercial interest. However, some material (poetry etc.) is seen to have long-term value. Cannot foresee a time when it will not have commercial value.

Rights holders/managers

***Publishers** keep an archive in case it is needed for further business etc. But they cannot be relied on as they are not doing it for the same reasons as an archive would.*

***Publishers** are archiving for the moment. But some CD-ROMs from the early days are gone, they have literally vanished.*

*There may be national strategy but keen that **institutions** can **protect their own materials**.*

*Feel that the **publishers have a responsibility** for maintaining the databases of electronic products. Each publisher should maintain own in-house archive. On national scale it would have to be sophisticated operation to accommodate range of systems and operating systems etc.*

***Data generators** might store it. But not every one, just the large ones or specialist ones. Specialist archives could specialise in different types of material. E.g. National Film Archive. Some data are being lost because no-one taking responsibility. Problem is ensuring complete coverage.*

Specialists and organisations with historical responsibility for archiving digital materials were also mentioned:

*Experimental data should continue to be handled by **specialist agencies***

***People who have historically done it** anyway. Any things held on other media are no different. Organisations which have preserved it in old media should translate their activities to new media*

Procedures for archiving

Legal deposit was seen by the majority of respondents as the most appropriate mechanism for ensuring that digital material is preserved. Further thinking is needed as to how this would work, although it was recognised that legal deposit would only apply to published materials.

***Legal deposit**. Deposit should not be voluntary.*

*One thinks of **legal deposit**. But what would it mean? There must be some legislative provision or else it wouldn't happen. Might be possible to get over fears of producers if **digital material were archived but not used for a number of years.***

*The BL is currently preparing the ground work for new legislation. Boston Spa has been negotiating with publishers about **legal deposit** of published materials and aims to arrive at an agreement with the CLA.*

*Digital material should be included in **legal deposit** legislation even though it is intangible. Many digital materials are genuinely new and should therefore be part of the national archive. However access to the documents by a network should be tightly controlled.*

*Hopeful for external **legal deposit** will build on journal publications - move from parallel publications to electronic only publications.*

*Similar situation as **legal deposit** of printed books. Other media should be treated in the same way.*

*Combination of carrot and stick approaches. Legal requirements for preservation of some information. **Legal deposit** for long term audit requirements. There are all kinds of medical records etc that need to be preserved for a long time to study disease demographics and other databases not covered by Public Record guidelines that ought to be protected.*

*Assuming something is the only version and tangible (e.g. CD-ROM) each **legal deposit** library should have a copy. If networked, access for all six. Publishers argue there should only be a single source as it is expensive to comply.*

The consensus seems to be that a voluntary code is unlikely to be satisfactory, although some people felt that a voluntary system would be better than nothing, until the legal deposit legislation is changed.

*Voluntary agreements between creators of data and repositories in UK? There is a good record for sound material, but everything else in the UK seems to work on a statutory basis. Legal deposit would seem to be the solution, but a **voluntary agreement could be tried** until then.*

*...refer to the music industry in relation to the National Sound Archive which operates on a **voluntary basis**.*

*Helpful to have extended legal deposit. In absence of that, we must **rely on voluntary agreements**.*

*Something akin to the existing national deposit requirements for printed material - i.e. it will be necessary to **rely on the publishers** to ensure that material is deposited.*

One respondent put forward the concept of Digital Object Identifiers (DOIs). These identifiers would incorporate existing labels such as ISBNs and could also include details of rights owners. ISWNs (International Standard Works Codes) could also be used in this context to identify publications, digital objects etc:

*Our view is that all digital works should have a Digital Object Identifier (DOI). This is more than an identifier; it is a receptacle for other numbering systems. For example, books get an ISBN; this is one number that could go into DOI, but an ISBN doesn't tell you about author and who owns rights as a DOI would. ISWC = International Standard Works Codes (a 'work' could be a book, digital object, film etc). ISWC could go into DOI so all objects are traceable. Where physically object is archived it does not matter. A **central register** would tell you that.*

The respondents referred specifically to national bodies or a central register to identify repositories and to pull together information on data archives:

***National body pulling it all together**. Tracking that data was deposited and logging it. This could be required at the point that the grant is awarded that it should be deposited with the relevant data archive. Could be extended to published material. It is all funded somewhere. Much archaeological stuff funded by developers. Archaeological curators give planning permission and this can stipulate the developers ensure the results are deposited in an archive.*

*Waste of effort to archive it twice. They would feel quite upset to know that the BL was archiving everything that they were doing. Where there is an existing data archive that is doing it well, pointless doing all the work twice. Could develop an **"Accredited data archive" system**. Leave it up to disciplines to decide what needed archiving. Quality judgements needed on what to archive and what not to. People working in the discipline are best placed to decide.*

*Some way of having a **central register** of the strengths of particular institutions. Way we can recognise that a particular repository holds the item. Trouble is that everyone wants to be on the register. Need to look at the technological expertise of the organisation and its ability to migrate the data. You could then have a number of data archives with responsibility in particular subject areas. Co-operation needed to decide what goes where. OK as long as data is freely available afterwards. This system could cause problems if the material is only available in one place physically because of copyright constraints.*

Another suggestion was for a national audit of digital materials that could be archived:

*May have to do something proactively like a **national audit**, at periods of time decided by relevant authorities. Information in categories called for deposit. Sampling exercise. Happy for the material to be held in libraries as long as* [they are available on] *single terminals not networked, similar to book access.*

Other alternatives suggested were: licence agreements, purchase, and economic reward:

*We are concentrating on published materials. They are considering different acquisition strategies including legal deposit, **purchase**, **licence agreements** to acquire electronic materials or the rights to use it.*

*Make it **economically rewarding** to deposit data. As people realise the value locked up in that data they will do it more.*

Timescale

Most of the respondents felt that selected material should be kept for ever. The key seems to be that once a digital object has been selected for preservation it is worth keeping for ever. The real issue is in selecting material for the digital archive and agreeing the criteria for selection:

*Like paper. Most material should be there **for ever**.*

*Most is held **in perpetuity.** There is some debate over whether all tide gauge readings, which are taken every five seconds, should be retained, or whether it would suffice to hold records for every 24 hours. Generally should apply same considerations as to paper records. Some material dates from the last century.*

*Generally, **in perpetuity**. Would depend on legal deposit. Factors that need to be taken into account: identifying national collection.*

*Holding material **in perpetuity** is an ideal. But this would involve reviewing the mode of access every five years or so. Unless material in different format.*

*If the material is part of the national record, then it must be kept **in perpetuity**.*

*Many people write for posterity, without hope of financial reward. Such material may deserve to be preserved. **No time limit** should be contemplated. The creative process is dynamic, cannot easily be contained within traditional archiving systems, and it is impossible to decide what should be preserved. The philosophy of preservation needs consideration; why preserve what; information as an entity, are but two questions. Whilst some material must be preserved, it would be wrong to impose control systems upon what should be kept.*

*Depends on material and how easy it is to archive. If a decision is made to preserve material then it is kept **in perpetuity.** Need to eliminate unnecessary duplication, although some is required for security.*

*Why not **for ever**? Any reasons why not would be trivial (e.g. storage space) -Would hate to think Doomsday book was chucked after 300 years because people had thought it was no longer of interest.*

***Ideally, no time limit**.*

***For ever** if the practicalities can be achieved. This principle is applied to paper publications and should be applied to digital ones.*

*Rights holder may want to destroy his older stuff. Should be preserved **for ever unless rights holder deems otherwise**.*

*The refereed journal literature should be preserved **in perpetuum.***

*No, **indefinitely** preserved*

***Keep for ever.** Whole point is you don't know what you want. Social historian can be interested in what could appear trivial.*

***Keep for ever.** As with any archive, you get into it never can say what is redundant. Not everything is of equal worth but can't tell in advance.*

***In perpetuity**. Important that they are unique items.*

*If you have made a judgement to archive it you should **never get rid of it**.*

*Their policy is once gone to trouble **preserve it for ever**. Are selective at start. Only way it can be used in future is if you have put resources into documenting etc. Selection up front. Don't own the material. Entered into a contract to preserve it for publishers so cant dispose. Identifying data sets to remove is more costly than refreshing.*

*If we take trouble to audit, decide on sampling should be should **preserve for ever**. It is a guideline to the state of scholarship at that time. Want to see what was available in that time.*

***No time limit**- there isn't a limit on printed material, and I don't believe digital material should be treated any differently.*

Some people saw a problem with managing this process and of providing access to digital material in the future:

*There should not be a black and white rule. Some materials have an infinite life and others have a limited life. There is a **high management cost** associated with a selective policy. It is impossible to predict the future usefulness of material. We are also looking at time sequence material, but it is impossible to get access to it while it is in use. **Costs of selectivity** and guidelines for selection versus costs of preservation of everything - also problem of locating material.*

*Don't know. Public records are kept for ever. **Original medium problem**. Means of moving to new media or preserving it and ability read it. In technical terms, so many rapid changes that a generation lasts no more than 3-5 years. More than 10-15 years preservation could be a problem. We need to build up an industry involved with archiving which can find a way to keep it long term.*

However some respondents felt that a time limit would be appropriate for certain materials:

*For certain products there **has to be a time limit**. Books are not locked into a technology. Now have applications which run on MSDOS. By the year 2000, finding a machine with DOS could be a problem. We need to turn to the IT industry for help in setting standards and documenting what equipment was needed to read which versions of software and operating systems.*

Other respondents felt there should be minimum retention periods or periodic reviews:

*There should be **no absolute rule.** Perhaps certain minima, e.g. for full duration of copyright while material has some value. Outer limit should be left to archivists.*

*We archive material for as long as it has **commercial value**. With these rights go responsibilities.*

Does depend how important it is (for example, if a cure is found for cancer, then Medline items on cancer will probably be of little importance in 200 years time).

*Some material should be **reviewed from time to time** to see if it is still important (e.g. archive of aerial photos).*

***Market forces** will determine whether things are reissued.*

*Note that the amount of material in storage will grow. May need to **discard some on a regular basis**.*

*If magnetic media costs keep falling can forget about disposal, but assuming they hit a level, may need a disposal policy. Not an expert. Disposal is a second go at collection policy. It should be seen as a retention policy rather than disposal. **Reappraise the collection**.*

Format of material

There were two groups of views on the format in which material should be kept:

1. a standard format
2. original format of the archived material

Many of the respondents expressed a desire for maintenance of digital materials in a standard format without specifying the format:

*Whatever is the **standard language** so that material can be reformatted over time. Would rely on advice of technical staff.*

*Format that meets current international standard if there is one. Should aim for **international preservation standard**, as was the case for microform. Access should be convenient.*

***Standards** are the crux of the matter. Would go with a standard if there was one. Impetus needed at government level to get started.*

*Ultimately need **standardised format** because technology is always changing. Either constantly reformat or reformat once and say we have particular format. Otherwise need a museum of obsolete technology to look at it.*

*Based on what I have seen, wherever possible digital materials should be **as standard as possible** for ease of migration later. However publishers may require materials to be in the published form which may not be standard.*

Specific formats suggested include:

*Important to adhere to standards such as **SGML** and **MARC**.*

***HTML** is helping because it is widely accepted. There is some conflict about the cataloguing of electronic material, and there have been calls for replacements for ISBN for the electronic world.*

*Relates to the uses. Works which need to be commercially accessible should be stored in **tagged format like SGML.***

*Or insist that final preserved version is usable based on **ASCII** or **simple building block format**.*

*It is now clear that the world wide web and html are so widely accepted, that if something new was developed, it would have to be made compatible with this. So we should make **HTML** the model.*

*There is a problem with the publishers. They want the archived material to look and feel exactly as it was created. Publishers and authors both have their reputations to think about. Publishers reputation includes the typesetting etc. The authors general perception is that being published on the web is not as prestigious as being published by a serious publisher. If their work appears as an **HTML** document, they could feel that their work is demeaned. This needs further exploration with author and publisher representatives.*

*If the choice is to have an archive in **HTML** or not at all, surely most would chose to have it.*

*We send material to users in image format **Acrobat PDF** (Postscript Document Format) or **Tiff** files (bit mapped images) no character based files are sent. Seems to be adequate at moment. People use different standards so want to present users with as few formats as possible, **limit choices** or better still, stick to one.*

*Impossible to hold as is. We **actively migrate it**. Standard transfer systems **ASCII** output that is readable by future database systems. Principle follows for all classes of data. You need to preserve functionality, (what you could do with it) rather than what it looked like. Sometimes what it looks like might be part of functionality e.g. image, but generally the format not important.*

*Ideal that must try for is to go for standards like **HTML STML JPEG and MPEG**. Not always possible because the software we use to reproduce it can't always support them.*

***Text format with rich metadata.** You can store images etc as text files. The metadata tells you what you are looking at and what you need to view it. (e.g. a certain processor or emulator - are there any computer studies looking at this kind of issue?)*

*Yes - follow Canadian pattern. The Canadians decided to ignore hypertext links and came up with some other solutions. Working with a small output of Canadian publishers all humanities. Straight text. Science and medicine text not whole message. Graphs etc are the message. No use just taking raw ASCII - lose half the message. Needs to be in **form that carries the whole message** e.g**. PDF, Catchword, HTML** and **Tiff files**.*

The reasons for keeping in the original format were expressed in a number of ways:

*If preserving pure history, we need to keep the **original medium**.*

*Depends on the material. Many people want the format as well as the information content. Therefore often needs to be **kept in original format**.*

*Format should be durable and cost effective. It should **not distort appearance or content** of material. Should preserve a complete and accurate representation of material archived.*

*This is the single biggest problem. In theory it would be nice to have the **original material, content and delivery** but he can't see how you could achieve that.*

*Why can't the material be archived in the **format it is published in.** Technology seems to solve the problems it creates quickly. Provided there is a demand, people will think of a way to read old technology.*

*We would push for preservation in **format in which the documents were created**. Information content is not all that is important; there is a lot of value in the context and way data is presented.*

*Whatever else you do you must **preserve the original bits in their unmutilated form.** May be appropriate to migrate into standard form if it is reasonably priced. But only migrate a copy, **keep the original form too**. We can't force publishers to deposit stuff in certain forms and unpublished it will be in whatever form the individual has. Preserve original copy sequence of bits. After transformations/migrations there will be loss of data.*

*If is an attempt is made to standardise format that itself will go out of date. **Take material in the form it was produced in**.*

Some respondents also addressed this question in terms of the physical media for storage of the digital material:

*Current format seems to be **CD-ROM**, but needs to be more easily unpackable. Need better industry standards.*

*Nowadays most people use **PCs** or **Apples**. Probably there is a gap in the records when Amstrad word processors were in common use.*

*Standards for **CD-ROM** are OK*

*Whatever state in when archived should always be put in **large storage medium like DVD-ROM**.*

Funding for archiving

A digital archive should be funded by the government. It should be a national responsibility. This was the most common response to the question of funding:

*Payment should be a **national responsibility**. On the one hand, our organisation [a repository] is partly government funded, and thus has a certain responsibility for its data. However, it would not want to spend additional money for digital archiving. There should be **central funding for a national archive**. We would not want to take responsibility for reformatting. Our data is collected in terms of projects. Most projects do not make provision for longer maintenance.*

*Should be **national expense**. Top-sliced from national organisation or institution responsible for archiving. National Heritage budget.*

***The people**. Perhaps there could be incentives such as tax breaks for any private sector involvement.*

*If national repository, then funding should be part of general funding, i.e. **state funded**. If it is an added responsibility, this should be recognised in the setting of funding. Libraries must make a case for funding. If funds are not forthcoming, then priorities will have to be adjusted.*

***Tax payer**. Could have a charging mechanism to access the data. This should be favourable to the academic community.*

Government responsibility was often linked to the British Library and other deposit libraries. Many people thought the system for print publications was suitable for digital archives:

> ***The government**. Libraries.*
>
> *Closely linked to the **current legal deposit**. There is a danger in having two different systems.*
>
> ***The government**. Organised in the same way as we have for **printed material**. It is part of the nation's heritage. There is no difference between digital and print. This will not happen unless the government pays.*
>
> ***Government** has a primary responsibility. Publishers should be expected to donate a requisite number of copies. Organised in the same way as at present for print. Principles don't change, just the media*
>
> *Has to be the **government**…*
>
> *Stick with **British Library** model. As the Library gets things for free, it should be prepared to do the warehousing. Need some kind of standard numbering system for labelling of digital documents to cut down on staff time spent cataloguing (cf. ISBN). Government departments should keep own internal records. They would also benefit from good labelling. If material has market value, then the market will pay for it.*
>
> ***The National Library**.*
>
> *The **Library** ...would be prepared to consider paying for [relevant] material if so required.*
>
> *If the digital material is part of the **national collection**, it should be funded in the same way as any other part. Payment is open to negotiation depending on the form of deposit. The organisation of this task would keep someone in a job for life.*
>
> ***Public domain funding** as in funding of **legal deposit libraries**. This should apply across the board - archival and published material. International - EU legislation? For funding, otherwise identify a predominant country. The LDLs have a claim on material originating abroad and sold substantially in the UK and Ireland. This needs to be projected positively.*
>
> ***Public domain**, but not much problem in getting private sponsorship to contribute. Should it be national or international? In UK we can organise it in a parochial way. Type of tax. Thinking of EEC as co-ordinating body. This would be preferable. Might have interesting ways of storing multilingual stuff. Nice to think that EU countries are all on same wavelength. Central funding. These decisions should be made internationally. Different storage policies are not helpful. National borders are increasingly irrelevant.*

We note that the two responses above included the notion of international collaboration. One person doubted that any government funding would be forthcoming:

> *The government has said there will be no additional resource. Long term archiving is a separate issue.*

A number of respondents felt that publishers or creators of information (e.g. academics) had some role to play in funding:

> ***Our organisation [a publisher] is prepared to pay** for the archiving of material so long as it is of commercial value. After that the archiving body can take responsibility.*
>
> *Has to be the government. **Publishers will pay to keep own archives**. But what if they go bust or out of business? At the moment we have got six deposit libraries. Publishers put their books in. Use of it by a scholar doesn't really affect the publisher's business even when the book is still in print. It is easier to buy it than to go to a deposit library. Slight problem with very expensive ones as people might be willing to travel to use these. In the future there will be a single digital archive.*

*The publisher. The **people who create the information** should put aside money for its preservation. Can associate a certain cost with preserving stuff. If we don't collect income from creators, we will not be able to preserve. Information has to become self supporting. This should be viable. Charge for access. Charge different kinds of people for different kinds of access. For example, academic users pay a little, pharmaceutical companies pay a lot. So complex to preserve and so expensive, we have to charge for access.*

Several people thought that funding should be shared among various sources, particularly if there were to be a distributed archive:

*Financing will come in many forms. Within the legal deposit scheme, the Dublin copy should be replaced by a digital copy to be retained in UK, and **publishers should contribute** to the cost of preservation.*

*Government is responsible for ensuring the framework is in place to ensure it happens. Originators and publishers need to contribute to maintenance of materials. **Top-slicing** or a tax to pay for this - a levy on materials which are produced or sold - like a national insurance system.*

***Whoever is storing it**. National listing solution. With globalisation it seems inappropriate for the UK to store only British works (which could be produced say in the States anyway). For academic purposes may be worth paying for a government funded archive. We could never apply to government to store stuff they will exploit commercially, nor could publishers.*

*Once the critical mass is reached, it will become clear that **scholars and scientists** should pool their resources to protect and perpetuate the corpus; the corpus, by the way, will have to be distributed and multiply redundant. The sponsors can be government (part of the research funding budget, perhaps), universities and research institutions (out of paper serial cancellation savings), and perhaps learned societies.*

*The DNH stresses that there are no extra funds available for this data archive. A **distributed archive would be cheaper** than a central one because many of the archives already exist. I am very dubious that the deposit libraries could make sufficient economies through more co-operation etc to enable them to fund a centralised digital archive. Like the Knowledge Gallery model, **fees could be charged for access** to the archive material (ideally for deposit as well, but I do not think you would get away with this!). It would still have to be **subsidised** however. It is not realistic to charge everyone enough to cover the full costs so you would need:*

A high rate for access by commercial organisations

Government subsidy

*The J-Store model is interesting here. They started off with several million dollars in foundation money with which they did the first digitising. Then they invited **academic libraries to become subscribers**. They charged a large up front sum to join and a much smaller annual fee to use the archive. The annual fee covers the cost of access. The initial fees were invested and the interest pays for maintenance and more digitising. They reached a point where it rolls and it seems to work. There is probably only room for one J-store in the world. The agreement of the publishers is essential. In return for allowing people to access their journals through the J-store, they get their back files digitised for them.*

*Partnership between libraries and funding bodies. Digital preservation will cost extra. Start up costs high as it is new. It is a national function therefore needs extra funding. Legal deposit libraries get extra funding. Need not be held in a library but still an extra cost. **Costs** should be **shared between libraries and host bodies** e.g. universities and producers as it is in their interest that it should survive.)*

*Analogy of print, **cost shared between publisher and library**. Same model could apply. Does not know if it is more expensive to store print or digital.*

***Users** of the material **should pay ultimately**. Nation should pay on behalf of its future users. For each library, it should be part of core task for which they receive funding. HE sector should pay for the bits it keeps. It is not practical to charge on point of access, it should be a collective system.*

We need to know more precisely what it will all cost. Impossible to get someone to agree to fund something if they don't know the scale. Tony Hendley did some costings, but by his own admission they were very rough and had a very wide range per item. We need more studies with different costing models based on different ways of doing it and based on actual projects. The BL is best placed to make these estimates. Some companies now sell storage space via the Internet. They could bid for the storage aspect, so the librarian's job would only be in selecting materials and format.

British Library thinking on digital materials is of interest here:

*The BL is currently preparing a proposal for funding an IT infrastructure to support electronic publications in the collection under the **Private Finance Initiative** (PFI). As part of this they will be looking at the future funding of electronic publications for posterity. We want to treat digital materials in the same way as printed publications. They should be subject to an acquisitions policy, retention and disposal policy, access, services based on the collection. They are making a case for extension of legal deposit to electronic documents - CD-ROMS have been one of the focuses for this, but want to extend this to electronic journals, Web sites and dynamic databases. Technology is needed for this but they cannot afford this with existing resources. Estimate £20M needed. The procurement would be based on PFI funding. We are making bids to other funding agencies focusing on the heritage aspect.*

Several respondents have mentioned in passing the possibility of charging for the use of an archive. One was more specific:

*Who pays? Depends on why material is being preserved. The **beneficiaries become the obvious source of funding**. Preservation of scholarly knowledge, commercial value, scientific value, defence against threats. We have devised a matrix which will be published of who benefits against what services offered in order to prioritise allocation of resources*

Organisation of funding

How should funding be organised? Many people had already answered this question, e.g. as is currently the case for paper, co-operative efforts, incentives such as tax breaks, top-slicing. Others provided additional thoughts:

*No idea but ideally the deal could be on a national level so it was **free at point of use**.*

*Either **each application for grant includes money for archiving**, a one off payment to the archive at point of deposit. Administratively expensive. Lots of little charges. Or a research funding body funding lots of projects, top slices it and pays the archive to preserve it. This happens now as grant comes from JISC. Funding bodies like English Heritage also could be looked to for **top sliced grants**.*

Depends on the source of the material. In private information the organisation is responsible. Public information, our cultural heritage, should be funded from the public purse. Difficult because we have suffered short termism in government.

***Same way as for printed publications**. Bearing in mind talking about audit which would not place an equal burden on all publishers.*

***Centrally funded body** in same way as for print. If nation decides it needs this for future generations and it will have to be kept long after publishers have ceased making money from selling it; it is unreasonable to expect them [the nation] to foot the cost.*

*Libraries own funding; don't see another option than for funding to come from the **repositories**. Not producers of the material. Certain amount of material can be voluntarily donated. A lot is like short print runs. Cost could be prohibitive. Publisher is not only providing expensive copy but undermining his ability to sell more copies. People will be reluctant to deposit for public use expensive stuff. This then means that only popular stuff is freely available.*

*The **same organisations who pay** for the national collections **now**. In the same way as it is organised now.*

If you look at publishers who makes journals available online, they have been unwilling to say how long we will give access to this. Not going to keep everything for ever - all wondering what to do. Not just a commercial thing. What happens after this? What do we say to people who have paid subscriptions and we no longer give access to the back issues? Answer is to pay by the drink. BL supply back issues and get a reasonable income from it. Pay ***copyright fees****. Part of the profit for each.*

Legal and commercial considerations

How to provide users with access to material while protecting the interests of copyright holders lies at the heart of digital archiving. Views on access ranged from the extremes of denying all access (i.e. material should be held purely for preservation) to making material freely available at no cost.

For most respondents, access should be allowed but with some restrictions. Some people thought that the same rules should apply as for access to paper archives. Others specified ways in which the differences between digital material and paper required a different approach to be taken.

Publishers are concerned that if publications in an archive are easy to consult they may be losing potential sales. This position is magnified if the publications can be obtained via digital networking. A related concern is that of downloading.

One person pointed out that standard copyright rules apply to electronic material. He added:

At the moment with legal deposit, any individual can look at archives and photocopy under fair dealing. With electronic publications, the publishers say it is impossible to fair deal and they are threatening to sue. They are much more concerned about depositing e-publications. Some publishers ... invest an enormous amount in a CD-ROM which will only sell about 10 copies at an enormous price, don't like the idea of it being on a network. They have a point. Similarly with high value financial information.

A commonly held view among respondents was that digitally archived material should only be consulted at a single stand-alone PC. Opinions differed as to whether this should be one PC in one place or one PC in each deposit library (or other repository body).

While some respondents wanted free access to all material, others were prepared to make some charges. For example:

Users should be allowed to browse without payment. If they want additional service, the charge should be passed on, as with photocopying.

Several people spoke of licensing arrangements:

Our products carry a user's ***licence agreement****. Such agreements should be thorough and fair and cover issues like fair dealing.*

One concern is people stealing stuff which is available over the Internet and then publishing it. … ***License fees for access*** *could be demanded. There will be an ongoing saga of sorting out funding arrangements.*

Issues to do with reproduction of material. Would have to be subject of discussion between rights holders and BL. ***Licensing arrangement*** *could be worked out.*

There are legal issues surrounding preservation, or which have an implication on preservation. For example, we need to look at the ***legal effects of the way licences run****. They are time bound, not perpetual. If you pay for an e-journal and they raise the fee and you stop subscribing, they take everything they have sent to you in the past; you have no back issues or discontinued runs.*

A number of people believed that no access (other than technical or curatorial) should be given whilst the publication was being exploited commercially by the publisher. Thus the material would be retained for purely preservation purposes to

begin with. A related view is that copyright holders should be recompensed while material is current. A respondent suggested that

> *There should be restrictions on use similar to the photocopy restrictions for printed materials.*

However, he warned

> *If digital archiving is publicly funded, then it is not practical to recompense publishers. This is a difficult area to legislate for.*

Some respondents thought that access should depend on the use to be made of the material. For example, it should be for scholarly rather than commercial use. One person thought that restrictions on downloading would be enough to preclude commercial use. Another said it was often difficult to distinguish between commercial and academic use.

A number of people felt that users should only have access to material if they could prove they needed to see it, and that users should have to make an effort. For example:

> *Anyone who can **demonstrate a need** to see them. BL reader's ticket is a good idea. Need to make a case for the need to see the material … they should have to demonstrate their need and make sure they cannot get it from their public library first.*
>
> *Users still need to be required to **make an effort** to gain access, prove that they have a bona fide reason etc. They should also be required to prove that they can't get hold of what they are looking for elsewhere. E.g. a student looking for an e-journal should be going through his university library who will have to pay.*
>
> *Anyone who can show **they can't get the information elsewhere**. Condition: When it is no longer available commercially, or out of copyright.*
>
> *In a legal deposit context **any bona fide user** who can **demonstrate need**. Serious research interest. Not available frivolously.*

Distinctions were made about the nature of the material. One respondent expressed the view:

> *The copyright concerns of trade and scholarly book and textbook and popular magazine authors and publishers are* completely *different from those of the authors of refereed learned serial articles; indeed, the latter are in a profound conflict of interest with the (paper) publishers of that learned serial corpus.*
>
> *Copyright exists to protect authors and publishers from two kinds of theft: (1) Theft of* text *(through contraband reading or selling) and theft of* authorship *(through plagiarism, etc).*
>
> *Trade authors and publishers have a common interest in protecting their products from both forms of theft. Both publishers and authors lose revenue if their joint product is stolen rather than fairly paid for. All the worse if they are re-sold under another's name.*
>
> *But the authors of refereed journal articles are concerned only about theft of authorship. For them, theft of their text is a victimless crime. It is precisely to be read by as many people as are interested that they publish their research in the first place. Here they are in a conflict of interest with their publishers, and this will have to be sorted out somehow.*
>
> *My prediction is that the trade model will have to be abandoned for the learned serial literature. The much reduced per-page cost of electronic-only journals will not be recovered through subscriptions, site licences, or pay-per-view (the three cost-recovery models paper publishers are currently considering) but by author page charges, covered either by the author's research grants, or the author's institution, or an outright subsidy by a consortium of interested parties (governments, research foundations, institutions of research and higher learning) in exchange for a learned serial corpus that is free for all, globally, and in perpetuum.*

I know it's the "perpetuum" that you are worried about, but it is too early to worry, and probably too early to try to impose standards; what is needed first is the accumulation of a sufficiently large mass of literature, and a readership (and authorship) that are sufficiently committed to USING the electronic corpus as the locus classicus of the learned research literature for them. so that they all have a vested interest in preserving what they are already so critically relying on now. That day, when necessity will spawn the requisite invention, cannot be pre-empted by orderly planning, etc.

Anarchic accumulation of our intellectual goods onto the new medium must take place first. By my lights, there will be two digital archives: trade and non-trade. My concern and expertise is only in the latter. Commercial interests are not involved there, and authors retain copyright (as there is no need or point in assigning it, given that they publish royalty-free and welcome the theft of their texts).

Many respondents consider that questions of access and copyright should be subject to negotiation. One person suggested a series of generic rules that could be adapted by negotiation in specific cases. Another introduced an argument for an expert committee which could restrict access altogether or for a period of time; it could ensure very limited access to some things (e.g. not put on a network, no printing, stand-alone PC). Another felt that access should be controlled by the place of deposit.

There was some mention of the need for IT mechanisms to be put in place to ensure that copying of material is not possible. A respondent referred to a French database of museum objects which has a system embedded in it to stop downloading and printing, and to the Electronic Copyright Management System (ECMS), which has produced watermarking (a mark is embedded into a document which will show on all printouts to identify the publisher, even if the text is edited).

Other security issues relate to user identification and payment systems, which are covered by different projects:

*JISC are developing an **authentication method** which is reliable, foolproof, simple method for people to prove that they are a student or bona fide user. Project called ATHENS. It can also keep a record of what a person has used/looked at. An aggregated version of this information can be sent to the publisher.*

*Work is also going on on **micropayment systems** where a very small amount (0.5p) is charged for every access. It would probably not be cost effective to gather this small amount of money.*

How do you stop people downloading more than they should?

*ECMS Electronic Copyright Management System. Projects like Imprimatur. Some of these projects have come up with **smart card systems**. Individuals have a card which can be put into the PC which identifies them and the use they are entitled to (i.e. how much they have paid). It identifies what you are allowed to do with this particular database and disables any functions such as printing or downloading if you haven't paid enough. They sent the most developed one to a university computer department and challenged the students to hack it. They couldn't.*

Suggestions for increasing security included adding something to the metadata (a lock or authentication mechanism of some kind), and 'simple encryption mechanisms to stop people "saving as" or using in a way you don't want them to; can put a lot of security into PDF files'.

Other legal considerations

When asked what other legal and commercial conditions applied to the preservation of digital archives, a common response concerned moral rights. For example:

*Should be understood that creators have in UK **moral rights**. They are:*

Integrity (not corrupting the work)

Paternity (right to be named as the author)

Misattribution (right not to be accused of being author of something you didn't write)

British authors may assert their rights. If they don't assert them, they don't have them. Not an unwaivable right as it is in some countries; a publisher may prevail on him to give them up. There may be an EU directive on this in due course. Most authors assert their moral rights. Not copyright but other rights (e.g. if you write an anti smoking article a tobacco company cannot repeat it sentence by sentence to argue against it) .

Issues surrounding legal deposit were also touched upon at this point:

*The **legal requirement to deposit** digital data should be comparable to that applied to paper records (e.g. material given to the Public Records Office). Note that some of our data is commercially confidential. We do some work for overseas governments, and cannot always release the data it produces to those other than the client. If there is a framework of centres responsible for national archiving, then each centre should use its discretion on such cases.*

***Legal deposit** repositories should not be given new powers to allow preservation. No copying of non-print material should be allowed for preservation. If a recording is damaged the repository should be entitled to request another or if not possible, to get specific consent to create another copy, perhaps from another source.*

Several people spoke of the importance of metadata. For example:

***Metadata** needs to be attached to all documents. Allows a way of representing copyright holder, who they are and attach different details if copyright changes. Copyright is different from moral rights and both are needed.*

This led on to software and agreements between publishers/distributors and software houses:

***Software** and **supporting data** is an issue. Sometimes the publisher is not the owner of the copyright for the software. There are agreements with third parties which may preclude deposit arrangements. The software producers may have to become involved and they are difficult to identify and may be outside the UK. We use Dataware who are based in the USA.*

*In support, it is not just the publisher involved, but **software house**. Publisher signs an agreement with software house who agree to support technical problems. Complex legal documentation. Term for the licence and publishers would not continue a licence for a product produced years ago if we no longer use that software. A library holding national archive could incur problems here as the software house has an interest. Licenses would have to be held centrally. What if software house stops trading etc.*

Some software houses agree to have their code held by a central body so if they go out of business their customers have the safety net of going to the central body. We could make this arrangement mandatory (at moment only concerned publishers who make sure it is written into the agreement).

Is *software a subject of this archive? It is a creative piece of work.*

Other concerns included: privacy/data protection, contract law, PR Act, authentication of material, liability, ownership of data, purging and definition of a publication:

*Deposit of material, **privacy**, maintaining **authenticity** (not subject to alteration), **intellectual integrity,** moral right of author, **defamation**. **Contract law** may override statutory privileges.*

***Data Protection**. This includes requirements for you to give access to individuals to data about them. May be legal issues about collection of data.*

*Archival authority has to carry forward all the constraints on the developer like **data protection**. These have to be honoured, continuity of contract.*

***PR Act** might have relevance. There are other Acts that might relate to information gathered by government. Collection use and retention of things like statistics. There are Acts around some items. One relates to agricultural statistics, income tax etc.*

*Other area is not tampering with it. **Authentication**. Once library receives an item, we have got to know that nobody can change it. May be that library has to prove that this is how it was published or when it was first available. Important with digital publications. The way to achieve this is to make a copy available for use and keep the original.*

*We need to be sure that it does not deprive publishers of sales, protect the **integrity of the publication**. Can't be seen to be repackaging publishers material for resale. We understand the nervousness of the publishers and want to make sure that there are no loopholes.*

*Our **liability** if we lose somebody's data. Hoping that unless can demonstrate negligence small print will hold up - they do not guarantee to preserve it for ever.*

***Ownership of data**, when distributing data abroad, if no reciprocal legal arrangement what do you do if a user of your archive makes information available on a web server in Iraq.*

*If it's decided to destroy [the publication] the archive copy could be the only copy. Before this happens there ought to be a question raised. A whole separate deliberation process on **purging** is needed.*

*BL ducked issue of **defining publication**. Said there was an equivalent to paper, but there is a very messy area of distribution. Much of the stuff they have in the archive is not defined as publications. No procedures for giving things ISBN numbers etc. Difficult to see how do that with Web sites. Can make changes to web sites so what is to be saved. What is the definitive version?*

FINDINGS OF THE FOCUS GROUPS

Four focus groups were held as part of the project, involving a total of 23 participants chosen to represent the views of:

Authors, Data Originators and Research Funders

Publishers

Distributors

Repositories

A list of all the participants is given as Appendix A to this report.

There was some overlap between groups because a few people were unable to attend 'their' event but agreed to participate on another occasion and, in any case, the distinctions between organisational roles were somewhat arbitrary. Several people would have been equally at home in two or more groups.

In each case, participants were asked to say whether their organisation had developed a policy or strategy for digital archiving and, if so, what form this took. They were then invited to participate in a structured brainstorming activity (based on the Nominal Group Technique) to arrive at priority issues for consideration and amplification during the remainder of the session. Participants were asked to:

- individually record their responses to a question about the main issues to be faced in ensuring that digital materials are preserved and are accessible
- contribute one idea each in turn to a master list, typically resulting in 20 to 40 topics
- select the five most important topics and rank these
- pool the results of this ranking exercise to show where the consensus and variations occurred
- discuss the high priority (combined rank orders) issues further.

Unsurprisingly, there were a number of common concerns across the four groups (such as achieving migration or emulation; copyright; and financial issues) but different priorities and emphases were also offered, as can be seen by comparing the highly ranked issues from each group (the words used are those offered in the groups):

Table 1 - Top ranked topics from the focus groups

Rank	Authors, Data Originators, Research Funders	Publishers	Distributors	Repositories
1	Common strategic approach by providers of preservation services (coherence, consistency and interoperability)	Preservation of the functionality of electronic publications	Criteria for permanent preservation/What is worth archiving?	Standards and common formats
2	Intellectual property rights	What constitutes a publication?	Contracts between archivists and information providers/ Co-operation between copyright holders to permit voluntary or legal deposit	Permanence and refreshment of data
3	Security (protecting against piracy; preservation against catastrophe; preservation of integrity -	Who should keep digital materials/ Rights' holders benefits /Who should pay?	Copyright issues	Initial capture into an electronic record-keeping system
4	Financial implications (Who pays? Who benefits?)	Need to use open standards in storing the data (e.g. SGML rather than MS Word)	Funding and costs	Access
5	Migration and emulation from one generation to the	Rights owners' benefits	How to organise all aspects of data emulation and migration	Financial viability and responsibility

Full lists of the topics offered by each group are given in Appendix D to this report.

Group overview

Authors, Data Originators and Research Funders

The Authors' group stressed the need for a common strategic approach to archiving digital materials and (unsurprisingly) for attention to intellectual property considerations, envisaging that the unilateral approach to decisions about preservation would have to give way to greater collaboration in future. They saw national co-ordination of digital archives as a more appropriate way forward than a centralised national digital repository, hoping to allow groups with specific interests to manage their own archives whilst ensuring that material was more generally available. They felt that a common approach to archiving digital material entailed providing a framework for guidelines (covering such issues as emulation and security as well as best practice in the area) which was not prescriptive, since otherwise it would be ignored.

They saw market demand as a major driving force in deciding what was to be archived. One suggestion was that a voluntary approach funded (at least initially)

by the interested parties would be more appropriate than asking for public funding for an unfocused national resource.

Publishers

The Publishers' group assumed that a national archiving strategy based on the Deposit Libraries would emerge. Accordingly, they saw a need to separate out the repository function of bodies such as the British Library and their role of providing a document delivery service to make information more widely available. The British Library was seen as a commercial document provider which was in effect competing on unfair terms with other peoples' materials.

They felt that transfer of electronic publications could take place at the point of publication *so long as there was an embargo on release to the general public*. If no guarantees could be made to publishers to protect their commercial interests the alternative rehearsed was for publishers to hand over their material at the point (defined by them) at which the publication was no longer of commercially exploitable value. So long as an electronic publication was commercially exploitable they felt that it was in the publishers' interests to maintain (and retain) it.

Although the group accepted the idea of preservation for the common good, they thought that the humanistic goal of national collections being maintained and preserved as a service to scholarship was no longer tenable because of the large volume of material and high preservation costs involved.

Distributors

The need to review the assumptions under which the deposit libraries operate was also taken up by this group. They saw the question of who should make the decision about whether or not to archive an item as requiring a collaboration between librarians, archivists, and publishers. Unlike the previous group, they felt that it was unrealistic to depend on the publishers alone to ensure preservation of digital publications, even in the short term.

This group saw preservation and access as separate issues, whilst recognising that preservation only has a purpose if someone has access to the material. However, they too accepted the idea of preservation for the public good, before extending this discussion into the area of museums preservation by arguing that different forms of technology should be preservable as well as exemplars of the technology:

> *CD-ROMs may survive for fifty or a hundred years but the software may not be available to interpret the data.*

Applying their broad historical perspective yet further, the group went on to argue that accidental preservation of material may continue to be significant because of the sheer volume of digital material that could be archived. They felt that there were various possibilities available from a systematic to an accidental approach. To some extent they felt that those items for which there is an on-going demand were likely to survive and perhaps that was good enough. (However, they recognised that The British Library would "*probably tend to favour some kind of deterministic approach*".)

Although some types of digital publication, such as CD-ROMs were seen as sufficiently coherent to be collected and archived, others, such as Internet sites were not. Hypertext links were seen as presenting particular problems because of the difficulty of preserving the ability to link with other sites.

Contracts between archivists and information providers, as well as co-operation amongst copyright holders to permit voluntary or legal deposit, were seen as part of the necessary response to the breaking down of traditional divisions of functions between publishers and repositories .

Overall, this group saw a key question as being whether the collecting strategy of the past centuries, based on keeping every edition, is still relevant today. Their view was that whereas in the past the printing process was sufficiently laborious to ensure that publishing was relatively rare, electronic material could change so quickly that this approach might no longer be appropriate

Repositories

Predictably, this group had a somewhat different view of access and deposit issues. They suggested that preservation and access in the future implies a requirement for an access copy and an archival copy of the data. The problem of legal deposit as it stands is that books on deposit are as accessible as any other book held by the British Library. However, for electronic documents the same principle probably should not apply. They saw the access and preservation roles becoming much more distinct for electronic documents.

Turning to large data sets, they felt that the source of the data would determine how these should be treated. Commercial publications should have strictly limited access with the emphasis on preservation. Access should be for preservation purposes only. Public data (such as that coming within the remit of the Public Record Office) should be more widely available.

There was felt to be a "heritage role for digital archives" and access by the academic community was seen as necessary. This was likely to lead to a number of new operating arrangements to secure appropriate access. The BFI was one example cited of an organisation that has based its policy on the access role rather than purely on preservation - partly enabled by the technology. The Data Archive has issued CD-ROMs of data for exclusive academic use and, where appropriate, proscribes use by commercially sponsored academics. This system depends on password access and undertakings by researchers not to use the data beyond agreed terms.

A single-user licence could limit access to networks, but difficulties would remain over downloading of data and its subsequent exploitation.

When rehearsing these general issues as well as in pursuing specific aspects of digital publishing (as distinct from the creation and maintenance of electronic databases and large data sets) there was a tendency for authors, distributors and other rights holders to find common cause and to express different concerns from the publishers and representatives of organisations concerned with archiving. These differences in perspective occurred at various points in the 'digital archiving life-cycle' as can be seen in Figure 2.

Figure 2: THE DIGITAL ARCHIVING PROCESS FOR PUBLICATIONS ISSUES RAISED IN FOCUS GROUPS

PUBLISHERS	RIGHTS HOLDERS	ARCHIVING BODIES	TECHNICAL ISSUES
POINT OF PUBLICATION			
Deposit now? Contracts with archivers?	Does deposit establish copyright? Co-operative approach?	Criteria for archiving? Deposit now? Networked access now?	Agreed standards. What constitutes a publication/edition?
Security: piracy; catastrophe; unauthorised changes	Maintaining rights after migration or emulation	Access by ...?	Preservation of functionality; open standards; migration and emulation
PRESUMED POINT OF DECLINE IN COMMERCIAL RETURN			
Deposit now? Allow networking by archives now?	Rights?	Exploitation licences? Access by ...?	Rights generated through transformation processes?
Ownership if resurgence of commercial interest	Rights if resurgence of commercial interest?	Maintenance charges to publishers?	
POINTS OF REVIEW FOR POSSIBLE DESTRUCTION			
Consulted?	Consulted?	Criteria for permanent preservation	The scale of the task

Some of these differences have already been reported above; the remainder are considered further below.

Key points from the group discussions

Some of the main points made in the group discussions are summarised under general headings below:

A common strategy and standards

General

The British Library role as an archiving organisation should be kept entirely separate from its role as a document delivery organisation. Indeed the British Library may not be the appropriate body, in which case we need to decide who manages the process and co-ordinates the work. It is partly a question of what Government is doing.

Any strategy must allow for the possibility that resources will be more limited in the future and that future archives will not be able to sustain the level of effort currently required for their management.

In addressing strategic issues, one participant suggested that there was a cultural difference between the US and the UK (and to an extent Europe) characterised by volunteerism in the US and state intervention in the UK.

JISC was felt to provide a useful model of how to proceed. It brings together national and specific interests and has a clear focus (on the academic community). The difficulty that the British Library or other national bodies face is that they are serving a much wider constituency with diverging interests.

There is a need for awareness in all organisations when creating electronic documents:

- a need to put a lot more into the migration plans
- recognition of specialist skills requirements
- recognition of the need to preserve material
- there is a need to educate electronic document creators

There are now collections that have been preserved for thirty years, hence there is experience to tap.

What is driving the need for archiving?

- preservation
- access and re-use
- commercial exploitation in the future

Standards

The role of standards is to enable interchange and exchange of information. Users need interoperability for better access to digital data.

Prescriptive standards in the electronic information world have failed to achieve full recognition. However the emphasis is now on 'permissive standards' such as SGML which does not tell document creators how to create the document or even what

software should be used, but does result in an environment that allows exchange of information. (Network Application Protocols to send information to other users formed another example cited of a permissive standard.)

Are common standards achievable? - Only if the IT industry moves towards interoperability. At the moment downwards compatibility is not assured. There are moves towards agreement to archive in common formats (ASCII, PostScript, and SGML) but this does not address database material. There is a need to identify standards which are most durable, to look at standards for objects or bundles of records. It is impossible to preserve formats of data so migration is important. There is no commercial advantage in trying to standardise.

Technical problems are not at issue. The concerns are more strategic in nature. There is a need to develop standards with the creators of data. Bodies such as AHDS are producing guidelines for creators. (Each of the service providers for AHDS has consulted widely with different stakeholders on this issue.)

Standards should be developed for a variety of different communities.

There was reported to be a lot of work going on this area. Brian Green and Mark Bide produced a report on *Unique Identifiers* which was published by the Book Information Council and revised in March 1997.

Broadcasting messages

Part of the dissemination process considered at the start of a project should include digital archiving of data and project outputs. In order to do this it will be necessary to gain the support of the research funding agencies.

There is a need to publicise the benefits of data preservation. However, real cost models are required.

Intellectual property rights

Copyright issues become complex if digital material is modified in any way, since it can be difficult to determine who has intellectual property rights invested in a document which has been processed (for instance during data emulation to maintain the information contained in a document) or amended by someone other than the original author.

It will be difficult to resolve many intellectual property issues until international legislation is introduced. The EC is publishing a White Paper in Autumn 1997 on *Fair use by libraries of electronic media*. The rights holders group felt that 'fair use' should not apply to electronic media because this interferes with commercial exploitation of data.

The copyright issue affects much more than just "books published on the Internet" and it is important not to lose sight of other types of digital material. For example, should copyright apply to transient data?

Practical steps such as the establishment of a database of rights holders could help to "keep track of the deep nesting of ownerships".

A distinction was offered between commercially-minded authors who would not like versions of their work to become widely available free of charge, and academics, whose interests would be met by their work being as widely distributed as possible. This distinction was immediately clouded by another view put forward that what drives most aspiring and actual novelists is immortality rather than money! It was

suggested that many of these people might be prepared to pay for their material to be preserved!

Financial implications

The level of commitment on the part of funding agencies is a key limiting factor for long-term preservation and needs to be taken into account in putting forward any recommendations. One possibility is that the level of resources that are devoted to digital archiving now will not be available in the future, and any strategy for long-term preservation must take this into account.

There is a need to determine the comparative costs of transfer or of emulation. Such estimates would provide a basis for the NPO or British Library to develop an integrated consistent strategy for archiving digital materials.

A more or less serious case was advanced for investigating the pornography industry, which possibly offers one of the most commercially successful areas of Internet commerce, where, typically, small access fees are offset by large numbers of transactions. The concerns about copying and re-distribution of images are put to one side because of the large volume of transactions.

One possible funding model is a nationally-funded system, such as the British Library model for printed books. Another approach would be to extend ISBNs to electronic publications and to charge a fee that covers future archiving costs. A third would be based on the principle that the user pays, with funds going to the archive depository or to the rights holders, or to both. Yet another possibility is to fund future archiving with a tax on current activities. The model suggested is that of the social security system where current contributions pay the pensions of retired people. (Unfortunately, this approach is susceptible to the same inherent flaw, that current providers will increasingly be outnumbered by 'dependants'.)

One group offered a series of observations on the value of archived material:

> *Immediate exploitation versus long-term keeping. The immediate exploitation gives tangible benefits and the long-term keeping gives intangible benefits.*
>
> *In some cases the longer the data is kept together the more likely it is to be financially exploitable.*
>
> *There is a progression from current, to current but not valued, to old and valued.*
>
> *The business case for keeping archives includes protection against litigation, maintenance of corporate identity and continuity. The heritage case is for historical and academic use.*

If the British Library or other organisation generates revenue for document supply, then some of this money should flow back to the rights holders.

In the view of the publishers' group, the handover point is when the commercial exploitation has come to an end - but this depends on how durable the commercial interest is likely to be.

Issues such as amount of usage, ease of access, cost of provision of access and of preservation need to be considered. These all help in the assessment of the commercial return. However there could be a problem if the level of usage of an electronic document resurges - does the publisher then get a return? If so, at what level?

One way around this would be for the publisher to retain ownership of the deposited materials and to get a royalty on accesses to electronic documents in the same way that royalties are paid for photocopies.

The caretaker approach would be to pay the British Library (or other archiving body) for keeping a document which is subsequently republished.

A further concern was over transfer of ownership if a publisher wished to sell a publication that was no longer of direct interest to its portfolio.

Co-operation is needed from all types of publisher. Non-profit publishers will not be able to afford to pay for archiving, so payment for archiving should be based on the ability to pay. Some publishers could pay a third party to archive material. Alternatively they could pay the British Library if they want to exploit their data in the future.

Another idea was to encourage funding by charitable trusts, as in the USA, or pump-priming to fund groups of publishers and repositories to set up a digital archive.

In the past academic institutions have paid for journals by subscribing to individual titles. Now as journals become available electronically (held nationally) national funders are negotiating block deals with publishers for the whole of academe in the UK to have access to electronic titles. National block funding for groups of users could be used as a model for archiving. (Central funding councils paying for block access.)

Decisions about archiving

Both of the currently available approaches to continued preservation, migration (or preserving the original data [object] and the platform necessary to interpret it) and emulation (where the information object is transferred into a new environment independently of the platform to ensure that it works in a similar way to the original) carry heavy operational costs, making selectivity in preservation essential. This in turn is informed by the question - why are we preserving material?

Selection of material is an issue - working on the basis that it would be too expensive to preserve everything. Who does the filtering? And who has the right to make decisions on what is kept? There is no consensus on whether there should be total or selective preservation. Some views suggest that selection criteria should be market-driven - this could be extended to usage of material. If archived material becomes corrupt or deteriorates because of lack of use it would not matter in a market driven system.

The system needs to be market driven. The publishers do not like the idea of top-slicing publication income to establish a national archive. The archive should be voluntary with publishers transferring material when it ceases to be of commercial value.

There is a danger of not making a decision about dealing with different types of information or sources that were formerly kept in paper form (such as rates payment records - a prime source for local historians). This could lead to breaks in existing series and loss of valuable historical material.

With book publishing there is a natural end-point, but with electronic publishing there is always the temptation for authors to go back and alter an existing text. This consideration suggests an accessions policy that freezes a publication at the point of acquisition.

One proposed definition of a publication is thatif an entity has an ISSN or ISBN, then it is a publication, although this does not address the issue of material on the Internet.

Some products are open-ended, such as bibliographic databases, so different criteria are needed for archiving. Snapshotsare one suggested solution to this. On-line discussion forums may require similar treatment with a round-up periodically and publication of summaries. However, there is no particular incentive for database providers to do this.

Three views on when to archive

It was argued by the publishers' group that rights owners should make the decision to publish and the British Library should make the decision on archiving but that the publisher must decide when it is appropriate to archive material. A definition of 'out of print' was needed for electronic publications, or a new criterion to determine at what point an electronic publication was handed over for archiving - the suggested definition was 'no longer commercially viable'.

The view of another group was that in devising a policy the principle should be that the originator is responsible for archiving a document when 'in print' but libraries are responsible when the publication is out of print. This means a transfer of responsibility between two bodies.

A strong case was made by the Repositories group for publishers to deposit electronic material as soon as it is published. The suggestion was that the British Library should get material under strict controls of access for an agreed period before was is made more widely accessible.

On a lighter note (?), a view emerged in two groups that perhaps accidental preservation is sufficient, since

after all this is the basis for most of recorded history!

Access

One participant advanced the view that deposit of material with an archive authority is useful because that authority is in a position to control access to the data and can therefore control its distribution much more effectively. For instance the authority could require that all users enter into a legally binding agreement not to breach the copyright constraints on use and distribution.

Academic users might be allowed privileged access to data through payment of a global license fee.

There is a need to consider the TV companies and the issue of access to old images. Small commercial TV companies come and go, leaving a problem in co-ordinating preservation. Video tape can be transferred to new tape technology, but it is very expensive and the future need is uncertain.

Archive management and preservation issues

Protection against catastrophe, involving back-ups of electronic media, off-site storage of back-ups, and a strategy for keeping materials up to date, necessarily adds to central archive costs, leading the Research Libraries Group in the United States (in their pilot study of preservation of digital materials at Yale) to conclude that preservation responsibilities should be distributed. However, the financial implications of preservation tend to lead to pressure for centralised storage because of economies of scale. The experience of such bodies as NERC which has had data archives for some time was thought to be particularly relevant.

The description of an object, changes in storage medium and encoding of documents are all changes that may need to be controlled or recorded in some way.

When a digital object is delivered to an archive there has to be a way of checking that what was sent is what is received. This process is time-consuming and expensive.

Preservation is currently based on perceived need. Different archiving groups are formed where there is a common interest in preserving a cluster of information - these are defined by the stakeholders. However there is no easy way of analysing cost benefits because these fluctuate as do the different groups involved.

Indexing is a key and neglected area. The software for indexing is not well developed. In moving material from one system to another the indexing functionality can be neglected or only executed to the fairly basic level of current CD-ROMs.

Core content may be preserved but there is a problem with links to other binary objects such as moving images or sound.

There is a need for long-term technical expertise in preservation teams to access the data in future and to interpret it correctly.

Why is digital archiving different from depositing a book? Electronic information is re-usable in a way in which books are not - it is inherently longer lasting.

Electronic publications present a problem because it is impossible to keep track of annotations and updates. It is possible to tell if a book has been tampered with.

The granularity of information will affect strategies. A book is a single entity, whereas an electronic document could be a record a paragraph or an entire document.

Two contentious views

Repositories might save pointers to the material and the publishers could maintain the actual archives. In that way the publishers could exercise some control on access.

Paper archives are preferable for parallel publications (print and electronic). Electronic documents with enriched content should be prioritised for digital preservation.

And finally a message for the ambitious: there is a person archiving the Internet who can be found at Architext.org

RECOMMENDATIONS

These recommendations arise from the findings of the investigation and are intended to provide a basis for further discussion and development of a national strategy for archiving digital materials. Although the remit of this study is focused on the responsibility for long-term preservation of digital materials, the recommendations must necessarily touch on other issues, such as standards and legislation that provide the framework for such a strategy.

Co-ordination

A body should be established to co-ordinate digital archiving activities in the public domain. This body (a suggested title is the National Office of Digital Archiving or NODA) would be responsible for implementing a national policy on digital archiving and on providing specialist input to the deliberations of the Library and Information Commission (responsible for advising government on information policy). NODA could evolve as an extension of an existing organisation such as the National Preservation Office.

The new body would provide a forum to represent the interests and views of the different stakeholders and it is essential that rights holders such as the producers of experimental data and publishers are represented on this forum.

The co-ordinating role should be separated from the archiving role which should be sub-contracted or delegated to specialist agencies. The national co-ordinating body would be responsible for development of appropriate standards of service (in consultation with the stakeholders) and for arbitrating where there are disputes about who should be responsible for which materials. The possibility of adopting a policing role as suggested for similar bodies in other countries should be considered as a way of ensuring that the archiving bodies are properly vetted.

Different approaches for different materials – a distributed archive

Actual archiving bodies contracted to keep national archives will depend on a number of factors, such as regional interest, type of data held, and ownership of material. For instance there may be a strong case for the National Library of Wales and the National Library of Scotland to keep digital material relating to Wales and Scotland. However certain specialist material such as sound recordings or experimental data may need to be kept by agencies with particular knowledge and understanding of data formats involved.

Arising directly from this approach would be the establishment of a national register of digital material that has been archived. The starting point for this would be a national audit of existing digital archives. Detailed catalogue records would be the responsibility of the individual archiving agencies, because of the widely varying nature of the digital materials that are archived. The feasibility of using a common descriptor such as a Digital Object Identifier (DOI) should be investigated.

We recommend that NODA work on a series of principles covering general categories of digital material rather than a prescriptive system with rigidly defined classes of data. This will allow flexibility and adaptability as circumstances change. Suggested categories include:

- Electronic publications including text, databases, digital sound and digital video
- Electronic public records – records generated electronically by public bodies in the course of their business and falling within the remit of the Public Record Acts
- Data generated in the course of publicly funded research or private-sector research regulated by national or European legislation (e.g. pharmaceutical research in support of product licences under the Medicines Act)

There is a recognition that publishers have different priorities from those of a national preservation body and for this reason it is unlikely that they would be appropriate organisations for digital archiving. However it is possible that some publishers with the expertise could bid to provide preservation services to the national body on a commercial basis.

Many publications and some experimental data fall across national boundaries. It is sometimes difficult to attribute a nationality to some of the larger publishing groups, especially when it comes to electronic publications which may be released in several different locations. NODA could provide a focus for liaison with other national bodies as well as transnational corporations and intergovernmental organisations.

Standards and guidelines

We recommend that NODA develops or co-ordinates the development of guidelines for retention and preservation of digital materials. Specific guidelines which apply to the creation of electronic documents or digital data should be developed so that individual items can be easily identified and managed in a digital archive. We suggest that NODA builds on the approach adopted by AHDS and other digital archiving agencies.

The choice of document format will have a significant impact on the ease of maintenance of digital archives. The use of standard formats or proprietary standards which are widespread will enable the data repositories to concentrate their resources on migration of materials from a few well-supported formats.

We recognise that in some instances there may be good reasons for preserving the original formats. In these cases we recommend that a parallel version is kept in a standard format to allow for migration to other supported formats in the future.

NODA should work with bodies such as the Public Record Office to develop guidelines directed at specific audiences, such as government departments. There is a lack of awareness of the need for electronic document management within the public sector and this needs to be made more widely known. The Public Record Office could develop guidelines on conservation and preservation of electronic documents that go beyond the current practice of printing them out and filing the paper copies.

Data generated in the course of research should be made available to the appropriate specialist repository. Funders should be encouraged to develop guidelines for researchers on the development of an archiving strategy for digital data. Some funders may choose to make a digital archiving strategy a mandatory part of any grant application. Deposit of experimental data whether publicly funded or not should remain voluntary.

Selection and permanent retention

Material that is selected for preservation should be kept for ever. The basis for selection should be the permanent value of the data or product. It is not possible to lay down rules that apply to all categories of digital material that may be subject to preservation. However general selection criteria should be established and should form the basis for the development of an archiving policy.

Our consultations suggest that the same principles used for selecting printed publications for retention could be applied to electronic publications. The responsibility for developing detailed acquisitions policies would lie with the legal deposit libraries or their agents.

Electronic documents in the public domain would fall under the Pubic Records Acts and policy should be established by the Public Record Office in consultation with the government departments and agencies originating the material. Materials that were deemed of historical value such as records relating to cabinet meetings would be preserved in their entirety. Other material relating to routine functions performed by individual departments might be sampled, to give future historians an insight into daily activities.

Experimental data and digital material arising from research present more of a problem for selection. Certain time-series data (e.g. population statistics) would probably need to be kept in its entirety for ever. For other data it may be sufficient to sample. For instance tidal readings taken at one minute intervals may not be of lasting value and hourly readings may be sufficient. The granularity of the data will be an important consideration for this type of data.

Databases present a particular problem because of their dynamic nature. A common archiving technique is to take a snapshot of the database at particular points in time. For some databases it may be necessary to keep an audit trail of all the changes made to a database over a period of time in order to get a comprehensive view of the data held. This may be the case for non-cumulative databases where individual records may be changed or deleted. It is not possible to establish a universal archiving policy for databases and decisions will probably have to be made on a case-by-case basis.

Although some of the people consulted put forward the idea of periodic reviews of archived material as part of a disposal policy we believe that this approach is incompatible with the philosophy of selecting material for permanent retention. It adds a significant and increasing overhead to the administration of the system and could lead to problems with changing priorities for retention.

Individual repositories should develop their own strategies to ensure the security and integrity of digital material. If necessary they should be allowed to make additional copies for preservation or conservation of digital data.

Funding

We recommend that digital archives in the public domain should be publicly funded. This is the only way to ensure continuity. However for electronic publications we recommend that publishers' contributions should be to provide one free copy of each electronically published title or issue that they produce. We suggest that copyright legislation governing legal deposit should be used. The cost of maintaining legal deposit material should fall within the existing arrangements for copyright materials. However in our opinion it will not be possible to comprehensively cover all electronic publications within existing resources. This

means that the legal deposit libraries will have to be very selective (and possibly reduce their acquisition of printed materials even more) or additional funds will be needed specifically for archiving electronic publications.

The cost of archiving experimental data should be subject to agreements to share costs between the research funders, the public, and the research communities that are likely to benefit from having access to materials in the long term. Public funding should be available to the depository agencies, though HE funding and via the research councils (or their successors). We recommend that users should not be required to pay at the point of access.

Legal deposit legislation

Electronic publications should be subject to legal deposit. It will be necessary to build into the legislation restrictions on the use of material. Clear definitions of what is an electronic publication would need to be established. Publishers should not have to deposit more than one copy nationally. We suggest that there are standard retention times for published material before they or the data on them are made available to researchers and the general public. Even after this time, certain restrictions on access may be required, such as a ban on networked access.

Decisions on the retention times before release to the public and other access restrictions should be made in consultation with the different interest groups including the rights holders, the depository agencies and the users.

BIBLIOGRAPHY

1. Arts and Humanities Data Service. *Digital Preservation.* Web page, last updated July 1997. (http://www.kcl.ac.uk/projects/ahds/background/other/preserve.htm)

2. *Copyright and the digital environment.* Managing Information 3 (1) Jan 96, p.25-6. ISSN: 13520229. (Statement prepared by the Library Association, UK /JCC working party on copyright which includes representatives of: Aslib, the Association for Information Management, the Institute of Information Scientists, the Standing Conference on National University Libraries, and the Society of Archivists.)

3. Department of National Heritage, Scottish Office, Welsh Office, Department of Education Northern Ireland. *Legal Deposit of Publications: a consultation paper.* Department of National Heritage, February 1997.

4. Hendley, A. *The Preservation of Digital Material.* London, British Library, 1996 (BL R&D Report 6242)

5. *Long Term Preservation of Electronic Materials. A JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib). Organised by UKOLN 27th and 28th November 1995 at the University of Warwick.* Report prepared by the Mark Fresko Consultancy. The British Library, 1996. BL R&D Report 6328. (http://ukoln.bath.ac.uk/fresko/warwick/intro.html)

6. Matthews, G, Poulter, A and Blagg, E. *Preservation of Digital Materials: Policy and Strategy for the UK. JISC/NPO Studies on the Preservation of Electronic Materials.* British Library Research and Innovation Centre, 1997. ISBN: 0-7123-3313-4, ISSN: 1366-8218. British Library Research and Innovation Report 41. (A version is available at http://lboro.ac.uk/departments/ls/staff/apoulter/digpres.html)

7. National Library of Australia, National Preservation Office. *Statement of Principles: Preservation of and Long-Term Access to Australian Digital Objects.*1997. (http://www.nla.gov.au/3/npo/natco/princ.html)

8. National Library of Australia. *Legal Deposit in Australia*, Fourth edition, 1997. Last updated 12 May 1997. (http://www.nla.gov.au/1/services/ldeposit.html)

9. National Library of Canada Electronic Publications Pilot Project. *Summary of the Final Report*, 1996. (http://www.collection.nlc-bnc.ca/e-coll-e/ereport.htm)

10. *Proposal for the Legal Deposit of Non-Print Publications: to the Department of National Heritage from the British Library*. January 1996.

11. Water, D and Garrett, J. *Preserving Digital Information.* Report of the Task Force on Archiving of Digital Information commissioned by The Commission on Preservation and Access and The Research Libraries Group, Inc. 1996. ISBN 1-887334-50-5. (http://www.rlg.org/ArchTF/)

APPENDIX A - PEOPLE CONSULTED

Focus Group Attendees

Anon., **a commercial management and academic publisher**

Mike Alexander, Document and Image Processing Manager, **BLDSC**

Susan Bennett, Archivist, **Royal Society of Arts**

Robert Bolick, **Stationery Office**

John Day, **Association of Learned and Professional Society Publishers, Biochemical Society**

Joy Foster, **Authors Licensing and Collection Society**

Dan Greenstein, Director, **Arts and Humanities Data Service Executive**

Peter Kibby, **TFPL Ltd**

Ed King, Collections and Preservation, **British Library**

Peter Leggate, Keeper of Scientific Books, **Radcliffe Science Library**

Trevor Lockwood, Secretary, **Author-Publisher Network**

Ian Macfarlane, Information Management Officer, **Public Record Office**

Clive Massey, **BIDS**

John McLaughlin, Director, **Association of Authors Agents**

Sally Morris, Director of Copyright and Licensing, **John Wiley and Sons Ltd**

Alan Morrison, **The Oxford Text Archive**

Keith Nettle, **Publishing consultant**

Elizabeth Ollard, **Humanities Research Board**

Chris Ostrom, Technical Director, **J. Whitaker and Sons Ltd**

Kelly Russell, eLib, **University of Warwick**

Ray Templeton, Librarian, **British Film Institute**

Judi Vernau, Director of Electronic Publishing, **Macmillan Ltd**

Bridget Winstanley, Director, Information and User Services, **The Data Archive**

Face-to-face Interviews

Neil Beagrie, **Arts and Humanities Data Service Executive**

Reg Carr, Bodley's Librarian, **Bodlean Library, University of Oxford**

Margaret Croucher, Research Analyst, **British Library Research and Innovation Centre**

Dan Greenstein, Director, **Arts and Humanities Data Service Executive**

Graham Matthews, Lecturer, **Department of Information and Library Studies, Loughborough University**

Charles Oppenheim, **International Institute for Electronic Library Research, de Montfort University**

Neil Smith, Network Services, **British Library Research and Innovation Centre**

Interviews by telephone and e-mail

Michael Dadd, Managing Director, **Biosis UK**

Andrew Baird, **British Library Document Supply Centre**

Ann Clarke, Assistant Director, Legal Deposit Project, **British Library**

Peter Cooper, Deputy Chief Executive, **The Royal Society**

Nigel Dickinson, **Dun and Bradstreet**

Robert Donaldson, **Learned Information (Europe) Ltd**

Lorraine Fannin, Director, **Scottish Publishers Association**

Lewis Flacks, Director of Legal Affairs, **International Federation of the Phonographic Industry**

Peter Fox, Librarian, **University of Cambridge**

John Goodier, **Goldhawk Information**

Stephen Harnad, Psychology Dept, **University of Southampton**

Iwan Jones, Conservation Officer, **National Library of Wales**

Denise Lievesley, Director, **The Data Archive**

Vanessa Marshall, Director, **National Preservation Office**

Ann Matheson, **National Library of Scotland**

Graham McKenna, Librarian, **British Geological Survey**

Colin Muid, **Cabinet Office**

Jim Parker, **Public Lending Rights**

Sean Philips, Librarian, **University College Dublin**

Julian Richards, **Archeology Data Service**

Seamus Ross, Director, **Humanities Computing and Information Management**

Chris Rushbridge, ELib Programme Director, **University of Warwick**

William Simpson, Librarian, **Trinity College Dublin**

Helen Smith, Director of Legal Affairs, **British Phonographic Industry**

Mike Snell, Library Development Manager, **Stirling University Library**

Sjoerd Vogt, Manager, **Knight Ridder Information Ltd**

Anthony Watkinson, **Thomson Science**

Robert Wellham, **Royal Society of Chemistry**

David Worlock, Chairman, **Electronic Publishing Services Ltd**

Alison Worthington, **Chadwyck-Healey**

Chris Zielinski, Secretary General, **Authors Licensing and Collection Society**

Interviewers and focus group facilitators

Monica Blake, **independent consultant**
David Haynes, **David Haynes Associates**
Tanya Jowett, **David Haynes Associates**
David Streatfield, **Information Management Associates**

APPENDIX B - SUMMARY OF FOCUS GROUP DISCUSSIONS

1. Focus Group for Authors, Data Originators and Research Funders

Tuesday, 24 June 1997
Isaac Newton Centre, London

Participants

Susan Bennett, **Royal Society of Arts**

Joy Foster, **Authors Licensing and Collection Society**

Dan Greenstein, **Arts and Humanities Data Service Executive**

David Haynes, **David Haynes Associates**

Trevor Lockwood, **Author Publisher Network**

John McLaughlin, **Association of Authors Agents**

Alan Morrison, **The Oxford Text Archive**

Elizabeth Ollard, **Humanities Research Board**

Kelly Russell, **eLib**

David Streatfield, **Information Management Associates**

Summary of conclusions

There were two main themes that emerged from the morning's discussions:

1. Common strategic approach to archiving digital materials
2. Intellectual property considerations

Common strategic approach

There is a weak argument for centralisation and intervention and a strong argument for co-ordination of a distributed resource. National co-ordination of digital archives is suggested as a more appropriate way forward than a centralised national digital archive. This will allow groups with specific interests to manage their own archives while ensuring that material is more generally available. A common approach to archiving digital material provides a framework for guidelines which are not prescriptive. They should include issues such as emulation and security as well as best practice in the area.

Market forces are a major driving force in deciding what is archived and what is not. One suggestion is that a voluntary approach funded (at least initially) by the interested parties would be more appropriate than asking for public funding for an unfocused national resource. Central intervention is considered ineffective in this instance, because it is difficult to focus on specific needs.

Intellectual property

Copyright issues become complex if digital material is modified in any way. It can be difficult to determine who has intellectual property rights invested in a document

which has been processed (for instance during data emulation to maintain the information contained in a document) or amended by someone other than the original author.

The level of commitment on the part of funding agencies is a key limiting factor for long-term preservation and need to be taken into account in putting forward any recommendations. One possibility is that the level of resources that are devoted to digital archiving now will not be available in the future, and any strategy for long-term preservation must take this into account.

There is a danger of not making a decision about dealing with different types of information or sources that were formerly kept in paper form. This could lead to breaks in existing series and loss of valuable historical material.

Prioritised topics

Scores for most favoured topics

1.	Common strategy...	17
2.	Intellectual property rights	16
3.	Security...	14
4.	Financial implications...	13
5.	Migration and emulation	12

1. Common strategy approach by providers of preservation services [Priority 1]

Coherence, consistency and interoperability.

It is not necessarily important that a common strategic approach is achieved. There is a need for a dynamic, fluid, and even anarchic approach to allow for change. Traditionally, prescriptive standards have imposed a constraint on the development of digital archives. However the role of standards has changed and the emphasis is now on describing what you have done. This message needs to be got across to users. SGML is an example of a permissive standards. It does not tell document creators how to create the document or even what software should be used to do so, but it does result in an environment that allows exchange of information. Network Application Protocols to send information to other users is another example of a permissive standard.

If everyone does his or her own thing it adds to the expense of maintaining archives. The role of standards is to enable interchange and exchange of information. Users need interoperability for better access to digital data.

Technical problems are not an issue. The concerns are more strategic in nature. There is a need to develop standards with the creators of data. Bodies such as AHDS is producing guidelines for creators. Each of the service providers for AHDS has consulted widely with different stakeholders on this issue.

Standards are not made widely available. They should be developed for a variety of different communities.

Part of the dissemination process considered at the start of a project should include digital archiving of data and project outputs. In order to do this it will be necessary to get the funding agencies on board.

3. Intellectual property rights [Priority 2]

The problem is that digital data, the "property", can be changed and republished easily. There is a lack of control of electronic resources in some environments (especially the Internet). It will be difficult to resolve many intellectual property issues until international legislation is sorted out. The EC is publishing a white paper in Autumn 1997 on "Fair use by libraries of electronic media". The rights holders feel that "fair use" should not apply to electronic media because it interferes with commercial exploitation of data. The European position is quite different from that adopted in the U.S.. ALCS feel that some sort of fair use has yet to be determined that can be applied to electronic media.

The copyright issue affects much more than just "books published on the Internet" and it is important not to lose sight of other types of digital material. Does copyright apply to transient data for instance.

One participant put forward the view that deposit of material with an archive authority has its advantages because that authority is in a position to control access to the data and can therefore control its distribution much more effectively. For instance it can require that all users enter into a legally binding agreement not to breach the copyright constraints on its use and distribution. This leads to the question of who you make the information accessible to: practical steps such as the establishment of a database of rights can help to keep track of the deep nesting of ownerships. It also provides the basis for a central clearing mechanism.

The other step is to appeal to the interests of the rights holders by the development of small consortia representing rights owners and academic users. The academic users are allowed privileged access to the data, possibly by payment of a global licence fee. There are some problem areas such as provision of access to back issues of journals. They may not be of current commercial interest to the publisher, but they may be of considerable value to current and future researchers.

There is a difference between authors who write for money and academics who do it for the recognition. Commercially minded authors will not want a version of his or her creation to become widely available free of charge whereas it is often in the interests of academics for their work to be as widely distributed as possible, so income is not so important a consideration. However, ALCS point out that although academics may not consider obtaining income for primary publication, they can and should share in downstream income from photocopying etc. Most novelists now earn more for secondary rights than for primary ones.

However another view put forward suggests that what drives novelists is immortality rather than money. There are an estimated 250,000 novelists or aspiring novelists in the UK. Many of these people may be prepared to pay for their material to be preserved.

Another view was that we should investigate the porn industry which possibly accounts for one of the commercially most successful areas of Internet commerce. Typically small access fees are offset by large numbers of transactions. The concerns about copying and re-distribution of images are put to one side because of the large volume of transactions.

The British Phonographic Industry (BPI) recently published some figures that suggested that every record is illegally copied six times.

The various parties have to have a vested interest in providing access to digital materials. Typically materials put on the Web are used for sales purposes and may

be loss-leaders, an initial chapter (a taster) or an abstract of the full item. The customer then has to approach the publisher or distributor to obtain the full item.

There is a need to investigate which economic models could be translated into an electronic arena.

4. Security [Priority 3]

Protecting against piracy and preservation against catastrophe and unauthorised changes.

Preservation of integrity. What is the authoritative text?

Protection against catastrophe is a managerial problem and includes back-ups of electronic media, off-site storage of back-ups, and a strategy for keeping materials up to date. The costs of protection are likely to be high and this would need to be taken into account in framing any legal deposit legislation. It is probably not viable to give individual creators the responsibility for archiving their own materials because of the high individual cost and the difficulty in policing this.

The Research Libraries Group in the United States did a pilot on preservation of digital materials at Yale. Their view is that preservation should be distributed. However the financial implications of preservation tend to lead to pressure for centralised storage because of economies of scale.

Other bodies such as NERC have had data archives for some time and their experience is useful in considering these issues.

Version control

Rather than going over the same ground it is important to look at existing standards and experience of others. For instance the problem of version control is addressed (at least in part) by the SGML standard.

Ann Kenny at Cornell University has done work on the manipulation of digital images and what is worth recording. With book publishing there is a natural end-point, but with electronic publishing there is always the temptation for authors to go back and alter an existing text. There needs to be an accessions policy that freezes a publication at the point of acquisition.

The description of an object, changes in storage medium and encoding of documents are all changes that may need to be controlled or recorded in some way.

CD-ROMs are considered more secure, because they cannot be changed once they have been pressed.

When a digital object is delivered to an archive there has to be a way of checking that what was sent is what was received. This is time-consuming and expensive.

6. Financial implications [Priority 4]

Who pays? Who benefits?

Archives already have a problem obtaining funding for long-term storage of digital materials. For this reason many institutions donate their archives to bodies such as research establishments that are better able to look after them.

One possible model is a nationally funded system, such as the BL model for printed books. Another approach would be to extend ISBNs to electronic publications and to charge a fee that covers future archiving costs.

User pays - with funds going to either the archive depository or to the rights holders or to both.

The costs will depend on what the archive has to do to keep a document viable over a long period. The commercial provider has a vested interest in preserving the integrity of their data. Outsourcing this is likely to be cheaper than doing it in-house. Exceptions are the very large publishers such as Reed-Elsevier which benefit from economies of scale.

Many printers keep magnetic tapes of typeset books (often without the author's permission).

Need to consider the TV companies and the issue of access to old images. Lots of small commercial TV companies go in and out of existence and there is a problem co-ordinating preservation.

There is a case for separating text-based materials from other types of material (video, images, sound). Need different strategies for archiving different types of material.

Many organisations may not be able to afford to preserve materials. However if they are of benefit to society, there is a case for public funding.

Need to publicise the benefits of data preservation. However real cost models are needed.

Selection of material is an issue - working on the basis that it would be too expensive to preserve everything. Who does the filtering? And who has the right to make decisions on what is kept? There is no consensus on whether there should be total or selective preservation. Some views suggest that selection criteria should be market-driven - let the market decide what should be kept by seeing what people are prepared to pay for. This could extend to usage of material. If archived material becomes corrupt or deteriorates because of lack of use it would not matter in a market driven system.

Preservation is based on need. Different archiving groups are formed where there is a common interest in preserving a cluster of information - these are defined by the stakeholders. However there is no easy way of analysing cost benefits because it changes and because of the different groups involved.

One participant suggested that there was a cultural difference between the US and the UK (and to an extent Europe). This is characterised by volunteerism in the US and state intervention in the UK.

JISC provides a useful model of how to proceed. It brings together national and specific interests and has a clear focus (on the academic community). The difficulty that the BL or other national bodies face is that they are serving a much wider constituency with diverging interests.

What is the purpose of preserving materials - this should be the driver for everything else.

Linkages

Would be additionally useful if users could link into other related material. The archive should innovate ways in which it delivers its services.

A matrix of stakeholders versus delivery modes can help to identify those cells where a revenue stream is likely. This provides the basis for prioritising particular areas of work.

A possible future is to fund future archiving with a tax on current activities. The model suggested is that of the old social security system where current contributions pay the pensions of retired people.

7. Migration and emulation from one generation to the next [Priority 5]

It is possible that resources will be more limited in the future and that future archives will not be able to sustain the level of effort currently required for managing archives. These anticipated limitations need to be accounted for in developing an archiving strategy

There are two possible approaches to preservation. The original data and the platform necessary to interpret it are preserved - i.e. preserve the platform and the object. The second possibility is emulation. The information object is transferred into a new environment independently of the platform (could affect copyright), to ensure that it works in a similar way to the original.

The cost of migrating information means that selectivity is needed. This in turn is informed by the question - why are we preserving material?

One suggested strategy is to make creators aware of the issues and of good practice.

Need to consider the future development of networks. Internet makes materials freely available. A more structured approach is needed in future. There may be charges for access to large parts of the Internet in future.

2. Focus Group for Publishers

Monday, 30 June 1997

Isaac Newton Centre, London

Participants

David Haynes, **David Haynes Associates**

Mr John Day, **Association of Learned and Professional Society Publishers**

Mr Keith Nettle, **Publishing consultant**

Peter Kibby, **TFPL Ltd**

Sally Morris, **John Wiley and Sons Ltd**

Chris Ostrom, **J. Whitaker and Sons Ltd**

David Streatfield, **Information Management Associates**

Summary of Conclusions

Transfer of electronic publications can take place at the point of publication so long as there is an embargo on release to the general public. The question then arises of how long the publication should be retained before release.

There is a need to separate out the repository function of bodies such as the British Library and their role of providing a document delivery service to make information more widely available. The British Library is a commercial document provider which is in effect competing on unfair terms with other peoples' materials.

If no guarantees could be made to publishers to protect their commercial interests the alternative is for publishers to hand over their material at the point (defined by them) at which the publication is no longer of commercially exploitable value.

The humanistic view of service to scholarship is untenable because of the large volume of material involved. This is not a viable way forward. The common good is not sufficient justification.

Preservation of materials - so long as an electronic publication is commercially exploitable it is in the publishers' interests to maintain the archive. When it is no longer the case, it is sensible to archive it.

The intent of making a document available electronically should provide the basis for selection of electronically archivable materials.

Prioritised topics

Scores for most heavily favoured topics:

4.	Preservation	8
2.	What constitutes a publication	7
6.	Responsibility	5
9.	Management of retrieval ..	5
11.	Who should pay	5

4. Preservation of the functionality of electronic publications [Priority 1]

Indexing is the key area and is neglected. The software for indexing is not well developed. In moving material from one system to another the indexing functionality can be neglected or only done to the fairly basic level of current CD-ROMs

Core content may be preserved by there is a problem with links to other binary objects such as moving images or sound.

The preservation of functionality has to cope with changing operating systems. This may be OK for the immediate future, but what about 1,000 years time.

What is the issue - preserving the content or the form? Preservation could be very expensive for the large volume of material being generated. It is the responsibility of data owners to preserve material.

The problem of preservation of TV programmes was raised. Video tape can be transferred to new tape technology, but it is very expensive and the future need is uncertain.

A view that emerged was that perhaps accidental preservation is sufficient, after all this is the basis for most of recorded history.

It is very difficult to set up a central store - distributed data is more realistic. Volume is not perceived as a problem but retrievability is an issue.

Need to determine the comparative costs of transfer or of emulation. This cost provides a basis for the NPO or BL to develop an integrated consistent strategy for archiving digital materials. There is still the issue of who decides what should be kept and how it is kept which needs to be addressed.

2. What constitutes a publication [Priority 2]

Although the scope of this consultation is wider than the archiving of electronic publications, the publishers' group were particularly interested in the role of electronic publications.

A lot of material is published but the owner must make the decision about what should be archived. However there was a countervailing view that it should NOT be down to the publisher to decide. The rights owner makes the decision to publisher and the BL can make the decision on archiving. However the publisher must decide when it is appropriate to archive material. A definition of 'out of print' is needed for electronic publications, or if that is not relevant a new criterion to determine at what point an electronic publication is handed over for archiving - suggested definition 'no longer commercially available'

The BL's dual role as a preserver/archiver and as a document delivery service was highlighted again.

The motivation of publications was also mentioned. Academics and politicians like to be published and perhaps have less vested interested in commercial exploitation of their publications.

It is increasingly likely that journals will not get published. What about hyperlinks?

Proposed Definition of a publication

If an entity has an ISSN or ISBN, then it is a publication, although this does not address the issue of material on the Internet.

Publications provide a basis for preservation of content.

There is a person archiving the Internet who can be found at Architext.org

One way of controlling access is by having membership subscriptions for access to certain materials.

There is a dual attitude to archives. Publishers do not like them and citizens do.

In devising a policy the principle is that the originator is responsible for archiving a document during the imprint but libraries are responsible when the publication is out of print. This means a transfer of responsibility between two bodies.

The CD-ROM SPAG was unhappy about the proposals put forward so far for digital archiving. One suggestion is that the archiving should focus on the words.

6. Who should keep digital materials / 8. Rights' holders benefits / 11. Who should pay? [Priority 3]

There is a bottleneck for publications and the problem is the need to process a publication to add value to it - limited human intervention.

Storage of everything is no longer realistic or affordable. Indeed is it desirable to keep everything?

Will the archive be usable? If it is who pays? The originators will have to pay initially for storage. There has to be a motivation for keeping the archives, otherwise it will not happen. However copyright legislation is too prescriptive and a tax burden. The system could become very bureaucratic rather like listed buildings.

One suggestion was that the repositories should save pointers to the material and the publishers should maintain the actual archives. In that way the publishers could exercise some control on access. Should access be limited to metadata on electronic publications - very expensive option.

The publishers should archive material until it ceases to be of commercial value.

The British Library caused ructions by requesting free access to electronic publications which (it is alleged) they were proposing to sell on as *.PDF files.

Originator to decide when a publication is to be archived. Equivalent to depositing a print publication when it is out of print.

1. To use open standards in storing the data (e.g. SGML rather than MS Word) [Priority 6]

Open standards are needed to achieve this - consistent identifiers - self-selecting group. Need to consider the level of granularity of metadata - it was suggested that this needs to be done down to article and chapter level.

Grey literature is difficult to deal with.

There is a lot of work going on this area. Brian Green and Mark Bide did a report on Unique Identifiers which was published by the Book Information Council and revised in March 1997.

8. Rights owners' benefits [Priority 7]

Authors have gained on publishers recently. Under CLA rules, a certain proportion of income goes to authors whose copyright has not been transferred to the publisher.

If the BL or other organisation generates revenue for document supply, then some of this money should flow back to the rights holders.

Publishers are concerned about handing over the rights to tamper with data. Databases 15 year rights apply.

The BL's role as an archiving organisation should be kept entirely separate from its role as a document delivery organisation. Indeed the BL may not be the appropriate body, in which case we need to decide who manages the process and co-ordinates the work. It is partly a question of what government is doing.

The system needs to be market driven. The publishers do not like the idea of top-slicing publication income to establish a national archive. The archive should be voluntary with publishers transferring material when it ceases to be of commercial value.

Continuously changing work

Older versions of documents are generally not of commercial value - customers pay for being up to date.

When a document is out of date the links are no longer of value - this gives a snapshot of part of the database that has changed.

The example of local authority rate books which are no longer used because the transactions are electronic highlights some of the problems of maintaining continuity and the integrity of historical series.

The users community has to ensure that its interests are properly represented.

One suggestion is that a paper archive is preferable for parallel publications (print and electronic). Electronic documents with enriched content should be prioritised for digital preservation.

Another suggestion is that maybe we need to archive less (e.g. resources such as the Internet contain a lot of junk).

The historical conditions that led to the development of libraries and museums do not exist now and therefore the development of new resources of this type (digital archives) is not possible on a national scale.

We need to ask ourselves where the money will come from and who will be looking after these resources for the next 1000 years.

There was a strong feeling that the whole process should be market driven.

3. Focus Group for Distributors

Wednesday, 2 July 1997
Isaac Newton Centre, London

Participants

Anon., **a commercial management and academic publisher**

David Haynes, **David Haynes Associates**

Ed King, Collections and Preservation, **British Library**

Clive Massey, **BIDS**

David Streatfield, **Information Management Associates**

Judi Vernau, **Macmillan Ltd**

General discussion

There was some general discussion on who would keep archive materials. Copyright implications were raised - a lot of what is being digitised will be re-used. Why is digital archiving different from depositing a book - electronic information is re-usable in a way in which books are not - inherently longer lasting.

Electronic publications present a problem because it is impossible to keep track of annotations and updates. It is possible to tell if a book has been tampered with.

The granularity of information will affect strategies. A book is a single entity, whereas an electronic document could be a record a paragraph or an entire document.

Books have an end-point. It is usually clear that a book is complete. Electronic documents need to be verified.

An archive would only take material from an authenticated source - is this true?

Databases present a particular problem because of continuous updating.

Lines can be drawn more clearly so that materials can have a date parameter and comments can be invited with a specific deadline.

Some things are open-ended such as bibliographic databases and so different criteria are needed for archiving - snapshots? Online discussion forums may require similar treatment with a round-up periodically and publication of summaries. Another way of describing this strategy is 'burying every 100th pot'. However there is no particular incentive for database providers to do this.

For academic journals a back run is often as important as the current issues, so that a user can search right back. There is therefore a commercial interest in keeping back-issues of journals.

There is a lot of material that does not need to be archived.

In the view of the publishers present, the handover point is when the commercial exploitation has come to an end - but this depends on how durable the commercial interest is likely to be.

An alternative would be to adopt a statutory approach to digital materials with extended legislation on legal deposit. This is not so difficult for CD-ROMs which

have a physical presence. However it is a lot harder for databases or other online resources.

This comes back to the problem of defining what is published.

The British Library is also interested in acquiring publishers' archives as well as archiving electronic publications.

There was further discussion about whether archiving should take place at the point of publication or at the point where a publication is no longer of public interest.

There is the problem of the BL having a dual role as preserver of information and as exploiter of information.

Issues such as amount of usage, ease of access, cost of provision of access and of preservation need to be considered. These all help in the assessment of the commercial return. However there could be a problem if the level of usage of an electronic document resurges - does the publisher then get a return? If so, at what level.

One way around this would be for the publisher to retain ownership of the deposited materials and to get a royalty on accesses to electronic documents in the same way that royalties are paid for photocopies.

The caretaker approach would be to pay the BL (or other archiving body) for keeping a document which is subsequently republished.

There is also the concern of transfer of ownership - if a publisher wishes to sell a publication that is no longer of direct interest to its portfolio.

One of the publishers felt strongly that the needs of information users were important. What does the library want in terms of digital archiving? This will affect what publishers priorities are.

Prioritised topics

Scores for most heavily favoured topics

Heading	Topic	Score
17.	Criteria for permanent preservation	8
4.	Contracts between archivists	5
11.	What is worth archiving	5
1.	Copyright issues	4
8 & 9.	Funding: & Costs	4
13.	How ...data emulation and migration	4
6.	Co-operation between copyright holders	3
10.	Accessible to whom?	3

17. Criteria for permanent preservation [linked to 11 - What is worth archiving?]

Non-copyright material may be different from purchased material which could be treated by copyright deposit law.

The question is who should made the decision about whether or not to archive an item, the archiving organisation or the publishers? It was suggested that it should

be a collaboration between librarians, archivists, and publishers. It is not realistic to depend on the publishers alone.

How will the information be used?

Preservation and access are separate issues. However preservation only has a point if someone has access to the material. Part of the reasoning behind this is 'for the public good'.

Forms of technology may be preservable as well as exemplars of the technology - this may be of interest, although this may be considered the preserve of a museum.

Accidental selection of material may be one way of coping with the large volume of digital material that could be archived. Essentially this could be a market-driven exercise - those items for which there is an on-going demand are likely to survive and perhaps that is good enough.

Need to consider form versus content.

Hypertext links present a problem because of the difficulty of preserving the ability to link with other sites.

4. Contracts between archivists and information providers [linked to 6 Co-operation between copyright holders to permit voluntary or legal deposit]

The traditional divisions of functions between publishers and repositories have been broken down.

There is too much potential material.

Some things are sufficiently coherent to be collected and archived e.g. CD-ROMs. Other things such as Internet sites are not.

Choice between a systematic approach and an accidental approach. The BL would probably tend to favour some kind of deterministic approach.

The solution to the issue of co-operation may depend on the criteria for selection.

There was the question of whether the collecting strategy of 100 years ago relevant today, i.e. keeping every edition applicable today. In the past the printing process was sufficiently laborious to ensure that publishing was relatively rare. Electronic material can change so quickly that this approach may no longer be appropriate.

1. Copyright issues

Copyright is the key to ownership and commercial exploitation of publications. This issue needs to be sorted out. Everything will fall into place as far as publishers are concerned. Journal articles will continue to be of ongoing interest. Textbooks translated into electronic media should be regarded as separate entities.

Dynamic entities need a strategy which is more like version control than archiving. For instance an encyclopaedia which is continuously updated and individual articles are archived as they are updated. At what point is there a definitive version of a document? This depends on future use (reference made to evidence -based healthcare at this point)

Archiving strategies: Audit of changes versus snapshots - how do you make the changes available? - This is a cataloguing problem.

2. Long-term expertise in preservation reams

Need technical expertise to access the data in future and to interpret it correctly. Long-term expertise is needed. It is not obvious when something has passed from being usable to non-usable.

8 & 9. Funding: profit and non-profit organisations and ability to contribute to archiving & Costs - expensive at present

Co-operation from all types of publisher - non-profit publishers will not be able to afford to pay for archiving - so should be on the basis of an ability to pay.

So publishers could pay a third party to archive material. Alternatively they could pay the BL if they want to exploit their data in the future.

The British Library's remit is to provide access to material that they hold, but we need to be aware of digitised data as well. The Digital Library, the British library's digitisation programme includes many items that are out of copyright.

Funding by charitable trusts, as in the USA, or pump-priming to fund groups of publishers and repositories to set up a digital archive.

In the past academic institutions have paid for journals by subscribing to individual titles. Now as journals become available electronically (held nationally) national funders are negotiating block deals with publishers for the whole of academe in the UK to have access to electronic titles. National block funding for groups of users could be used as a model for archiving. Central funding councils pay for block access.

13. How to organise all aspects of data emulation and migration

Financial implications of this.

Content versus delivery - if you emulate the interaction it will be expensive. However preservation of content is relatively straightforward.

In the medium term physical preservation is not a major issue, but the software is a problem especially from version to version. There is also a rights issue in emulation systems. It would be necessary to work out the relationship between the publisher and the BL.

4. Focus Group for Repositories

Friday, 4 July 1997
Isaac Newton Centre, London

Participants

Robert Bolick, **The Stationery Office**

David Haynes, **David Haynes Associates**

Peter Leggate, Keeper of Scientific Books, **Radcliffe Science Library**

Ian Macfarlane, **Public Record Office**

David Streatfield, **Information Management Associates**

Ray Templeton, **British Film Institute**

Bridget Winstanley, **The Data Archive**

Background and general discussion

The initial discussion focused on the definition of the scope of this investigation. Scanned images of existing publications are excluded from the scope of the project, but what about data that is re-keyed rather than scanned?

Prioritised topics

Scores for most heavily favoured topics

15.	Standards and common formats [linked to 1 & 27]	12
21.	Permanence and refreshment of data	6
2.	Initial capture into an electronic record-keeping system	5
12.	Access	5
26.	Financial viability and responsibility	5

15. Standards and common formats/Interoperability [linked to 1 & 27]

Are common standards achievable? - Only if IT industry moves towards interoperability. At the moment downwards compatibility is not assured. There are moves towards agreement to archive in common formats: ASCII, PostScript, and SGML, but this does not address database material. Need to identify standards which are most durable. Look at standards for objects or bundles of records. It is impossible to preserve formats of data so migration is important. There is no commercial advantage in trying to standardise. Is there pressure on suppliers to use common formats?

If someone assumes responsibility once the data is in the archive it can be controlled.

De facto standards include Adobe's *.PDF format. *.PDF files are generated using the Acrobat software. However there may be a need to conform to SGML. Adobe is converging towards *.XML as an Internet standard.

CD-ROMs may survive 50 or 100 years but the software may not be available to interpret the data.

What is kept? The physical record or the data?

Common formats - there is no format which can act as a universal panacea. Backward compatibility is desirable, if it could be enforced. E.g. ISO 9000 depends on backwards compatibility. Do we need to convert from one format to another?

Importance of specialist organisations doing long term preservation was emphasised. They can migrate the data to new formats.

Need awareness of all organisations when creating electronic documents:

- need to put a lot more into the migration plans
- recognition of specialist skills required
- recognition of need to preserve material
- need a policy and to educate creators

There are now collections that have been preserved for 30 years.

What is driving the need for archiving?

- preservation
- access and re-use
- commercial exploitation in the future

21. Permanence and refreshment of data

This is part of the previous discussion, but with emphasis on the technical issues. The issues are fairly clear. It is important that someone has taken on the responsibility for data preservation.

2. Initial capture into an electronic record-keeping system

Turning a record into a corporate record means documenting it properly. For electronic records this ties in with metadata. This can turn up in an electronic archive - the process of getting material onto the archive.

There is a need to educate users about the disciplines of generating electronic documents - including filing etc.

The Legal Evidence Act (1996?) states that all material on a case must be kept in one folder. This applies to electronic data. In practice, this means that e-mails are printed out but it does allow for electronic documents and e-mails to go onto electronic case folders. However it is important to ensure that there is a proper descriptive title.

This is about persuading people to understand what has to be done to ensure electronic documents are preserved.

A Canadian survey found that 15 seconds was the maximum amount of time that could be imposed on an author to index a newly-created document. Any more than that and records were either not indexed, or were done very quickly and often inaccurately. This process effectively has to be automated if it is to work in practice.

Promotion and training are key factors in this.

12. Access

Access into the future implies an access copy and an archival copy of the data.

The problem of legal deposit is that books on legal deposit are as accessible as any other book (held by the British Library). However for electronic documents the same principle probably should not apply. Compared with the book access and preservation roles will become much more separate for electronic documents.

The source of the data will also determine how it should be treated. Commercial publications should have strictly limited access with the emphasis on preservation. Access should be for preservation purposes only. Public data (such as that coming within the remit of the Public Record Office should be more widely available.

There is also a heritage role for digital archives and access to the academic community is probably appropriate in this case.

The BFI has based its policy on the access role rather than purely preservation - partly enabled by the technology.

Networkable products are a problem if only stand-alone access is provided.

The Data Archive has put out CD-ROMs of data for exclusive academic use - it does not allow use by commercially sponsored academics. This depends on password access and undertakings by researchers not to use the data beyond these terms or to pass them on to a third party.

For networks a single-user licence can limit access, but there is no control over downloading of data and its subsequent exploitation.

26. Financial viability and responsibility

Immediate exploitation versus long-term keeping. The immediate exploitation gives tangible benefits and the long-term keeping gives intangible benefits.

In some cases the longer the data is kept together the more likely it is to be financially exploitable.

There is a progression from current, to current but not valued, to old and valued.

The business case for keeping archives includes protection against litigation, maintenance of corporate identity and continuity. The heritage case is for historical and academic use.

If there is a fee for access, the ‘keeper’ gets some income, but so should the rights owner.

If material is kept for the national heritage, the publisher contributes the books and the BL keeps and maintains the items.

A strong case was made for publishers depositing electronic material as soon as it is published. The suggestion is that the BL gets material with strict controls of access for an agreed period before it is made more widely accessible.

There was further discussion about the definition of published material in an electronic context even though it was recognised that this study encompasses both published and unpublished material.

APPENDIX C - QUESTIONNAIRE

[Text in italics is for the guidance of the interviewer and should not be read out unless appropriate. In some cases it may be necessary to depart from the schedule where there are additional areas and issues that are relevant to this study]

Introduction

My name is *[name]* from David Haynes Associates. Is it still convenient to talk to you now? *[If not, reschedule and notify Tanya, or say "My colleague Tanya Jowett will contact you to try to arrange another appointment"]*

We are conducting these interviews for the National Preservation Office which was set up in 1984 to provide an independent focus for ensuring the preservation and continued accessibility of library and archive material held in the UK and Ireland. It has embarked on a major new initiative to develop a national strategy for preservation, digitisation and digital archiving on behalf of libraries and archives in the UK and Ireland. It recently established the Digital Archiving Working Group which is overseeing a number of projects including this one.

We are interested in your views about where the responsibility lies for the long-term preservation of digital material.

*[If they need a definition of digital archive material use the following: '**Electronically published information' or 'digital data' in the context of this project is data or text or documents originated in electronic form. This data may be made available to the public in electronic or printed form or it may result from publicly-funded research or be produced by public bodies. The scope does not cover conventional printed publications that are subsequently digitised for study or conservation. Compilations of data on CD-ROM, electronic databases and documents available via the Internet are included in this definition. Stationary and moving images and sound recordings held on digital media are also included in the scope of this investigation.**']*

But first I would like to ask you about your own role.

Background

1. Can you give a **brief** description of your job?

1.1 *[Unless already answered]* Does formulation of policy on digital materials form a significant part of your job?

Digital archiving policy

2. Does your organisation currently have a policy or strategy for dealing with the preservation of digital material?

2.1 If yes, what are the main areas covered?

2.2 Is there an available version of this *[policy document/strategy/plan]*?

Responsibility for preservation

3. In your view who *[which organisation]* should be responsible for long-term preservation of digital material? *[If they need a prompt list some of the stakeholders such as funders of research, authors, publishers, data distributors, or national repositories]*

3.1 What procedures should be put in place to ensure that the data is deposited with the appropriate organisation?

Timescale

4. Should there be a time limit for the preservation of digital archives?

4.1 If so, how long?

4.2 What factors need to be taken into account in determining how long to keep digital materials?

4.3 Why *[should there be a time limit]*?

Format of material

5. Do you have any views about the format in which digital material should be kept?

Money

6. Who should pay for the long-term preservation of digital materials?

6.1 How should this be organised?

Legal and commercial

7. How should the interests of the copyright holders be represented?

7.1 What other legal and commercial conditions apply to the preservation of digital archives?

Access

8. Who should have access to digital archives?

8.1 Under what conditions should they have access to the archives? *[If prompted say "For example should they pay at the point of access?"]*

Other issues

9. Are there any other aspects of digital archiving that we should be aware of?

9.1 If yes, what?

Other sources

10.1 Are there any other organisations or individuals that can throw light on this issue?

10.2 Are there any publications or other documents that you think may shed some light on this area?

10.3 If so, who *[please collect as full contact details as possible and include notes on who referred them and why]* or which publications *[full bibliographic details including a note on where to obtain it/them]*

APPENDIX D - INITIAL LISTS OF TOPICS OFFERED IN THE FOCUS GROUPS

Authors, Data Originators and Research Funders

- ♣ Common strategy approach by providers of preservation services
- ♣ Coherence, consistency and interoperability
- ♣ Overall adoption of agreed standards
- ♣ Metadata and version control
- ♣ Intellectual property rights
- ♣ Security
- ♣ Protecting against piracy and preservation against catastrophe and unauthorised changes.
- ♣ Preservation of integrity. What is the authoritative text?
- ♣ Fluid form access
- ♣ Accessibility. Ensuring that there are many different ways of getting at the data.
- ♣ Financial implications
- ♣ Who pays? Who benefits?
- ♣ Migration and emulation from one generation to the next
- ♣ How long material can/should be kept
- ♣ Concentrating on data creators
- ♣ Guidance on good practice
- ♣ Easy to use - remaining usable
- ♣ Legal deposit legislation
- ♣ Technology
- ♣ Changes in the technology
- ♣ Some discrimination - what to keep
- ♣ Rights holders' interests - technical and legal basis
- ♣ Who should actually keep the material
- ♣ Incentive
- ♣ Compulsory or voluntary?
- ♣ Access/distribution

Distributors

- ♠ Copyright issue - fair recompense to publishers (payment in perpetuity)
- ♠ Long term expertise in preservation reams
- ♠ Software continuity and media continuity
- ♠ Contracts between archivists and information providers
- ♠ Technology - how to handle further down the line : publishers supply applications and data?
- ♠ Co-operation between copyright holders to permit voluntary or statutory deposit
- ♠ Cataloguing issues
- ♠ Funding: profit versus non-profit organisations - ability to contribute to archiving
- ♠ Costs - expensive at present
- ♠ Accessible to whom?
- ♠ What is worth archiving and how to determine?

- ♠ Internet publications keeping track of unregulated publications (ISSN/ISBN plus...)
- ♠ How to organise all aspects of data emulation and migration
- ♠ Living archives for material published continually
- ♠ World-wide archiving
- ♠ Timescale - how long to keep the information?
- ♠ Criteria for permanent preservation

Publishers

- ♥ To use open standards in storing the data (e.g. SGML rather than MS Word)
- ♥ What is a publications/edition
- ♥ How do you deal with continuously changing content?
- ♥ Preservation of the functionality of electronic publications
- ♥ Long terms reliability of storage media
- ♥ Who should keep digital materials
- ♥ Rights holders benefits
- ♥ Management of retrieval and access
- ♥ What should be archived
- ♥ Who should pay
- ♥ What is the definition of out of print
- ♥ Ongoing commercial value of material
- ♥ Recovery of material from out of date..
- ♥ Obligations: originators versus libraries etc.
- ♥ Integrity of archived material
- ♥ How long material should be kept for
- ♥ Disaster recovery
- ♥ Foreign material

Repositories

- ♦ Common formats
- ♦ Initial capture into an electronic record-keeping system
- ♦ Metadata (formats, data on categories)
- ♦ Rapid obsolescence of software and hardware
- ♦ Security and protection
- ♦ Who should pay for keeping material?
- ♦ Structure of the archive (information retrieval or database system - object-oriented, relational or hierarchical)
- ♦ Maintenance as active archive by periodic migration
- ♦ Ownership
- ♦ Everyone is a publisher - no quality control of publications on the Web
- ♦ Authority and quality
- ♦ Access
- ♦ Authenticity of attribution by users (legal admissibility)
- ♦ What should be included - scope e.g. electronic mall
- ♦ Standards and common formats
- ♦ Organisational continuity of custody
- ♦ Added value - ability of users to do things with digital archives which is not possible with paper archives

- Who should be responsible for keeping the materials?
- The protocol/ease of capture from multiple sources
- Availability of suitable IT industry products
- Permanence and refreshment of data
- Information loss when migrating
- Constantly changing data
- Version control
- Retention periods
- Financial viability and responsibility
- Formats and inter-converting formats
- Dynamic databases
- Ensuring ownership gets value
- Should database archiving allow interactivity?
- Sensitivity/data protection
- Proselytising/promoting idea of digital archiving
- Internet charging
- Who and where? Which organisations should be responsible for digital archiving?
- Training
- Multi-media and dynamic documents
- Virtual indexes - pointers and links
- Interaction of subsequently digitised and printed materials with these.

APPENDIX E – POSSIBLE MODEL FOR DIGITAL ARCHIVING IN THE UK (PUT UP AS A WEB PAGE)

Digital Archives: who keeps them and who pays?

We'd like your views on this important issue.

Contents

Background

The United Kingdom's National Preservation Office (NPO) has commissioned David Haynes Associates to investigate attitudes towards responsibility for the preservation of digital data. This investigation is funded by JISC through the eLib Programme. It is associated with the NPO, and managed by the British Library Research and Innovation Centre, BLRIC . The Digital Archiving Working Group which is overseeing the studies wants feedback on who should be responsible for preserving digital data and electronic publications. We have started the consultation process by interviewing stakeholders such as publishers, authors, researchers, research funders, data compilers, archives, libraries and repositories.

Key questions

The two main questions that we are attempting to address are:
Who controls digital archives?
Who pays the cost of establishing and maintaining the digital archives?

Who controls the digital archives?

Rights Owners

If the rights holders (publishers, research funders, authors and performers, distributors) should be in control of the digital archives, how is continuity assured? For instance, if a publisher goes out of business or is subject to a take-over, or simply changes its areas of interest what guarantees are in place to ensure that the digital material continues to be maintained? Would rights holders be prepared to take on this responsibility?

Depositories

If a depository body takes on the responsibility for archive material a hand-over point has to be established. For instance, electronic publications could be handed over at the point of publication, or could be handed over at a pre-determined point (after a fixed period, or perhaps, after it ceases to be worth exploiting commercially). In this situation, it may be necessary to amend the legal deposit legislation to take into account digital materials. Under these circumstances would rights holders be prepared to pass on this responsibility?
Various suggestions have been made about which bodies should be responsible for looking after digital materials including:

National Preservation Office
Deposit libraries
National libraries
Higher Education (HE) institutions
Digital Archiving Authority

Who pays?
In our consultations three potential sources of funding for digital archiving were identified:

Rights owners
The rights owners include publishers and research funders. If they control the digital archives, then it is simplest if they are also responsible for paying for them. They would continue to have exclusive control of the rights of access and the right to continue exploiting the material commercially.

Public funds
Public funding for the preservation of digital materials would mean that users would not be charged at the point of use. The resources for a national digital archive would come from central government, either via existing agencies or through a new agency.

Users
Users could be charged for access to digital materials in one of two ways. They can pay at the point of access (pay per view), in which case the charges for popular items would have to be set at a level to cover the costs of preserving rarely used materials. The alternative is for the institutions to which the users are affiliated to pay a licence fee to cover use of digital materials, similar to the approach used for photocopying licenses granted to Higher Education institutions by the Copyright Licensing Agency.

Operating model

In order to help focus the discussion about Responsibility for Digital Archiving we have developed an operating model. This model represents one of the possible ways forward for digital archiving policy in the UK:
Distributed archive
Extension of legal deposit legislation to cover electronic publications
Specialist agencies for specialist material
Permanent retention
Selectivity
Standard formats for preservation
Public funding
Restricted access to publications for a fixed period

Digital data in the context of this project is data or text or documents originated in electronic form. This data may be made available to the public in electronic or printed form or it may result from publicly-funded research or be produced by public bodies. The scope does not cover conventional printed publications that are subsequently digitised for study or conservation. Compilations of data on CD-ROM, electronic databases and documents available via the Internet are included in this

definition. Stationary and moving images and sound recordings held on digital media are also included in the scope of this investigation.

Distributed Archive
A distributed network of agencies should share the responsibility for long term preservation of digital data. The activities of these agencies would be co-ordinated by a National Office of Digital Archiving (NODA) which would allocate responsibility for preserving materials in consultation with interested parties.

Extension of legal deposit legislation
Legal deposit legislation should be extended to include electronic publications. Other legislation such as the Public Record Acts may also need to be modified to take into account digital materials.

Specialist agencies
Some agencies will specialise by subject or medium such as Archaeology or sound recordings (for example the Arts and Humanities Data Service or the Data Archive), others will be more geographically selective (e.g. the National Library of Scotland for materials relating to Scotland).

Permanent retention
Once selected, digital archives should be kept for ever, because it is impossible to anticipate future needs. Historians often need data in context; discarding material on the basis that it is not being used could be a mistake. However this approach puts a great deal of responsibility onto the shoulders of those who do the selecting.

Selectivity
The volume of digital material generated is too large to consider archiving it all. Guidelines on selection criteria should be established and regularly reviewed by a central co-ordinating body.

Standard formats for preservation
As far as possible, digital materials should be stored in a standard format. Suggested standards include SGML, HTML, ASCII and Postscript for text documents, and D1 and D3 standards JPEG, and MPEG formats for film, video and audio material, for example. This allows for migration to new technology and reduces dependence on specific hardware and operating systems.
Many publishers feel that original formats should be preserved because the look of the publication, as well as its information content, convey a lot to the reader. One suggestion was that, whilst for preservation purposes it is necessary to convert the publication to a standard format, the original format should also be preserved so that data archaeologists can explore these in the future).

Restricted access to electronic publications
Archived material should be accessible under restricted conditions. There should be a standard retention time before material is made available to the public to allow for the commercial exploitation of the work.

Public funding
The system would be publicly funded through the British Library and legal deposit libraries in a similar way to the way in which legal deposit of printed material is funded (i.e. the publishers contribute by supplying a copy). Users should not be

required to pay at the point of access. Grants for research could be top-sliced to cover the archiving costs of the data arising from the research.

Your Views

We would like your views on:

1. **Who should be responsible for archiving digital materials?**
2. **How this should be paid for?**
3. **Whether the suggested** operating model **(described above) would in your view work.**

If you have a view about this model and want to join the debate please let us know. We would welcome your comments by 5 September 1997, although we will forward any comments arriving after that date to the Digital Archiving Working Group.

We can be contacted via e-mail or by contacting the address below:

David Haynes Associates

Signet House

49-51 Farringdon Road

London EC1M 3JB

Tel. 0171 242 4849

Fax. 0171 242 4858

Email:DHaynes1@compuserve.com